Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par
Université Toulouse III - Paul Sabatier
Discipline ou spécialité :
Ecologie

Présentée et soutenue par *Christine Lauzeral*
Le *20 septembre 2012*

Titre : *Prédiction du potentiel d'invasion des espèces non natives par des modèles de niche : approches méthodologiques et applications aux poissons d'eau douce sur le territoire français*

JURY

Dr. Núria Bonada (Université de Barcelone - Espagne)
Prof. Emili Garcia-Berthou (Université de Girone - Espagne)
Dr. Bernard Hugueny (Muséum National d'Histoire Naturelle - Paris)
Dr. Nicolas Poulet (ONEMA)
Dr. Pablo Tedesco (Muséum National d'Histoire Naturelle - Paris)

Ecole doctorale : *Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)*
Unité de recherche : *Laboratoire Evolution et Diversité Biologique (UMR 5174)*
Directeur(s) de Thèse : *Prof. Sébastien Brosse (Université Paul Sabatier - Toulouse III)*
Rapporteurs : *Prof. David Mouillot (Université Montpellier II)*
Dr. Didier Pont (IRSTEA - Antony)

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par
Université Toulouse III - Paul Sabatier
Discipline ou spécialité :
Ecologie

Présentée et soutenue par *Christine Lauzeral*
Le *20 septembre 2012*

Titre : *Prédiction du potentiel d'invasion des espèces non natives par des modèles de niche : approches méthodologiques et applications aux poissons d'eau douce sur le territoire français*

JURY

Dr. Núria Bonada (Université de Barcelone - Espagne)
Prof. Emili Garcia-Berthou (Université de Girone - Espagne)
Dr. Bernard Hugueny (Muséum National d'Histoire Naturelle - Paris)
Dr. Nicolas Poulet (ONEMA)
Dr. Pablo Tedesco (Muséum National d'Histoire Naturelle - Paris)

Ecole doctorale : *Sciences Ecologiques, Vétérinaires, Agronomiques et Bioingénieries (SEVAB)*
Unité de recherche : *Laboratoire Evolution et Diversité Biologique (UMR 5174)*
Directeur(s) de Thèse : *Prof. Sébastien Brosse (Université Paul Sabatier - Toulouse III)*
Rapporteurs : *Prof. David Mouillot (Université Montpellier II)*
Dr. Didier Pont (IRSTEA - Antony)

Remerciements

Avant tout, merci à Sébastien qui m'a fait confiance dès le M2 et m'a prise en stage puis en thèse malgré ma charge d'enseignement. Je le remercie en particulier pour la liberté qu'il m'a accordée tant pour le mémoire de M2 qu'ensuite pendant ma thèse. Un grand merci également pour m'avoir permis de passer deux semaines inoubliables en Guyane. Bref, merci pour tout, CHEF !

Merci à Gaël. Ses connaissances des modèles de niche m'ont beaucoup aidée. Et son implication dans le programme Refresh m'a donné l'occasion de découvrir des coins perdus du piedmont pyrénéen et de tester la résistance de mon foie. Merci aussi de m'avoir motivée pour me remettre à la photo.

Merci à Didier Pont et David Mouillot d'avoir accepté d'être rapporteurs de ma thèse et à Nuria Bonada, Emili Garcia-Berthou, Bernard Hugueny, Nicolas Poulet et Pablo Tedesco d'avoir fait parti**e** de mon jury.

Merci à Géraldine et Simon de m'avoir fait découvrir les vallées du Viaur et du Célé sous un éclairage nouveau. L'association pêche électrique – croustibon est inoubliable ! Et grâce à eux, je suis devenue le Lucky Luke de la soudure à l'étain.

Mes remerciements vont à Nicolas pour m'avoir fait confiance pour INVAQUA et à l'ONEMA. Sans leur travail de fourmis pour la collecte de données, une large part de mon travail de thèse n'aurait pas pu être faite. Et Merci à Flamby et Alex pour leur implication dans INVAQUA.

Une mention spéciale à Christophe et Jean-Baptiste qui m'ont fait redécouvrir les joies du Meccano et des mathématiques expliquées aux biologistes.

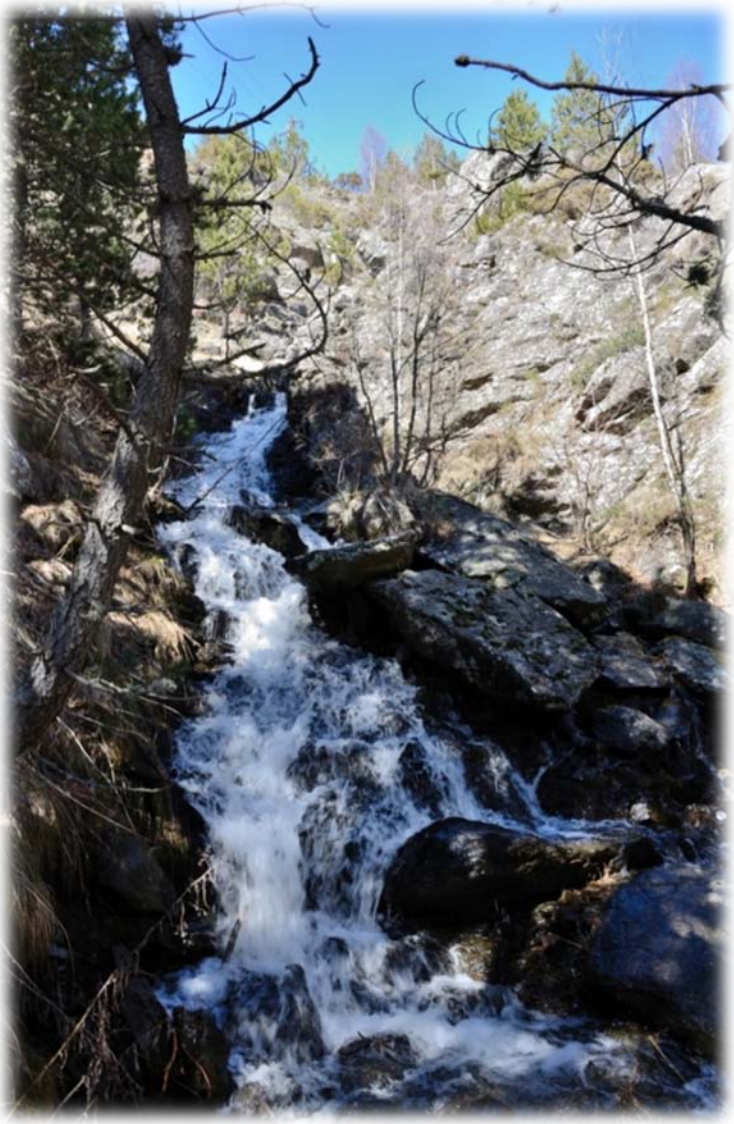
Merci à toute l'équipe AQUAECO pour son accueil**,** son endurance face à mes expériences culinaires et pour continuer à m'héberger en attendant que je parte vers d'autres cieux. Merci en particulier à Nicolas et Seb 2 d'avoir contribué à agrandir mon zoo.

Grâce à Pierre, j'ai enfin débuté la plongée. Je lui en sais gré.

Merci à Juliette et Aurélie qui ont fait preuve d'une très grande patien**c**e en particulier lors de la dernière année de ma thèse. Elles ont supporté sans faillir une maman peu disponible et parfois « légèrement stressée », voire vaguement « disjonctée ». Merci à Corinne et à Leslie pour leur amitié et leur soutien malgré leurs nombreux soucis.

Merci pour finir à Ben, le souvenir de nos instants de complicité m'a aidé à tenir les jours où le moral n'était pas au beau fixe.

Sommaire



Introduction

Le terme de changement global couvre toutes les modifications d'origine naturelle ou anthropique susceptibles de modifier la capacité de la Terre à héberger la vie (U.S. Global Change Research Act of 1990). Si la composante la plus connue du changement global est le changement climatique, de nombreux autres facteurs interviennent également. Parmi eux, les invasions biologiques sont une menace considérable pour l'ensemble des écosystèmes de la planète (Vitousek et al. 1997). Il apparaît indispensable de développer des outils adaptés permettant leur contrôle, d'autant que leur nombre et leurs impacts risquent d'augmenter sous l'effet du changement climatique (Rahel and Olden 2008). La lutte contre les invasions passe entre autres par la détermination des zones d'établissement potentiel des espèces à risque (Thuiller et al. 2005b). Mais les outils statistiques classiquement utilisés pour prédire la distribution des espèces rencontrent dans ce cadre des difficultés liées à la fois à la nature invasive de l'espèce (Gallien et al. 2012) et aux incertitudes concernant les conditions climatiques futures (IPCC 2007) et réclament donc des améliorations afin d'augmenter leur efficacité dans le cadre de la prévention des invasions.

A. Les changements climatiques

La planète a déjà connu par le passé des épisodes de changements climatiques brutaux, avec en particulier quatre épisodes glaciaires liés à des phénomènes astronomiques (oscillation de l'orbite terrestre, cycle de l'activité solaire…) sur les 400 000 dernières années (Figure 1). Ces variations du climat ont eu des impacts importants sur la faune et la flore. La dernière glaciation, en chassant l'homme de Neandertal d'Europe, aurait en particulier permis l'invasion de l'Europe par l'homme moderne (Müller et al. 2011). Ce dernier, par le développement de ses activités industrielles et la consommation massive d'énergies fossiles, est maintenant devenu un moteur majeur des variations du climat terrestre, en particulier en modifiant les concentrations en gaz et en poussières dans l'atmosphère.

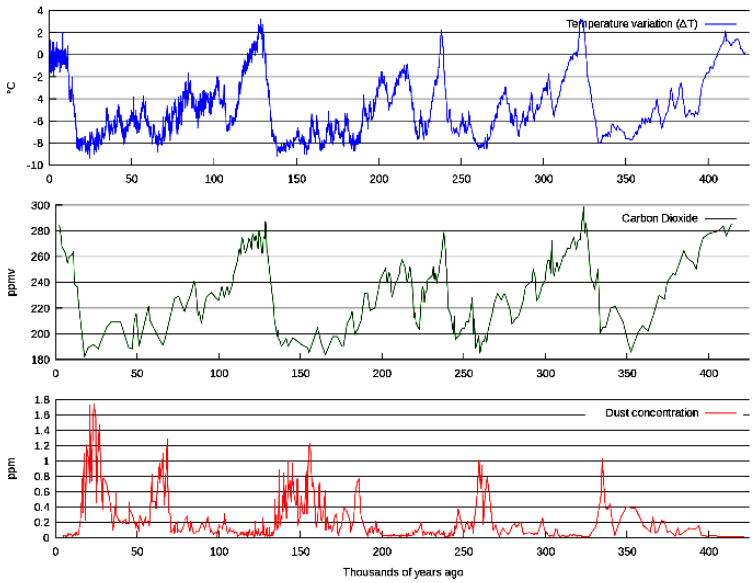


Figure 1 : Changements de température, de concentration en CO2 et de concentration en poussières dans l'atmosphère au cours des 400 000 dernières années. Les températures sont déduites des variations des concentrations en isotopes radioactifs (en bleu). Les concentrations en CO2 (vert) et en poussières (rouge) proviennent d'un carottage de glace en Antarctique (figure de William M. Connolley d'après les données de la National Oceanic and Atmospheric Administration, U.S. Department of Commerce, Paleoclimatology branch, Vostok Ice Core Data).

1. Définitions

Selon l'IPCC (International Panel on Climate Change), le changement climatique peut être défini comme « une variation de l'état du climat que l'on peut déceler (par exemple au moyen de tests statistiques) par des modifications de la moyenne et/ou de la variabilité de ses propriétés et qui persiste pendant une longue période, généralement pendant des décennies ou plus. ». Il concerne donc aussi bien les changements du climat dus à la variabilité naturelle que ceux liés à l'activité humaine. Cette définition diffère de celle de la FCCC (Framework Convention on Climate Change, également appelé GIEC en France) qui ne prend en compte que les changements d'origine anthropique : les changements climatiques sont les « changements de climat qui sont attribués directement ou indirectement à une activité humaine altérant la composition de l'atmosphère mondiale et qui viennent s'ajouter à la variabilité naturelle du climat observée au cours de périodes comparables ». Le changement

climatique est une des composantes majeures des changements globaux, c'est-à-dire de tous les changements dans l'environnement (climat, utilisation des sols, ressources en eau, chimie atmosphérique, systèmes écologiques) qui peuvent modifier la capacité de la terre à maintenir la vie (U.S. Global Change Research Act of 1990). Il est susceptible d'affecter tous les écosystèmes, mais aussi d'amplifier l'impact d'autres changements tels que la fragmentation des milieux et l'utilisation des sols.

2. Les observations

Les observations du climat sur les 100 dernières années mettent en évidence une augmentation de la température moyenne de l'ordre de 0.8° C.

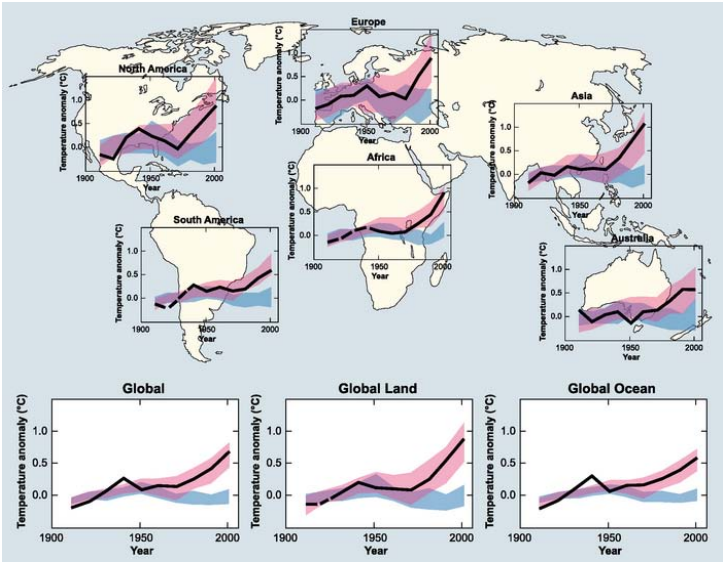


Figure 2 : Comparaison, à l'échelle continentale et à l'échelle du globe, des changements observés de la température de surface avec les résultats obtenus par les modèles climatiques avec un forçage radiatif naturel ou un forçage radiatif à la fois naturel et anthropogénique. Les anomalies des moyennes décadaires des observations sont montrées pour la période 1906-2005 et représentées au centre de la décade (ligne noire) et mesurées par rapport à la moyenne de la période 1901-1950. Les bandes bleues correspondent à l'intervalle 5%-95% de 19 simulations utilisant uniquement le forçage naturel (activité solaire, volcans). Les zones roses correspondent à l'intervalle 5%-95% de 58 simulations de 14 modèles utilisant les forçages naturel et anthropogénique (IPCC 2007).

Cette tendance n'est pas homogène sur l'ensemble du globe (Figure 2), avec une augmentation des températures plus importante pour l'hémisphère nord, l'Asie connaissant le

réchauffement le plus important (1.3°), alors que le réchauffement est plus réduit pour la surface océanique et les continents de l'hémisphère sud (0.7°).

Mais les changements climatiques n'affectent pas que la température. On a pu observer des modifications des précipitations (Figure 3) selon des patrons différents suivant les régions ainsi qu'une augmentation de la fréquence des évènements extrêmes (inondations, cyclones,…) et une élévation du niveau moyen de la mer.

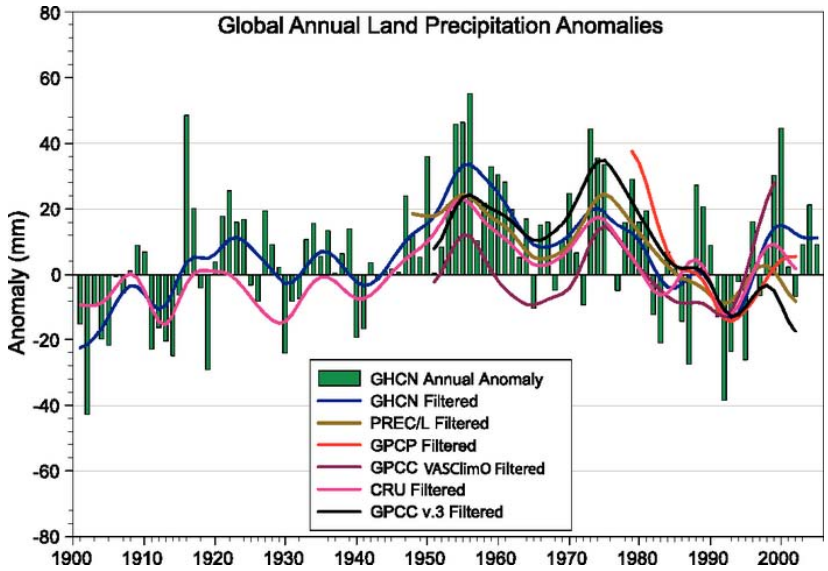


Figure 3 : Anomalies des précipitations annuelles totales (en mm) du GHCN sur la période 1900-2005 par rapport à la période 1981-2000. Les courbes correspondent aux variations décadaires pour 5 bases de données différentes.

3. Les prédictions

Il est maintenant largement admis que les changements climatiques observés ces dernières décennies sont essentiellement d'origine anthropique : « Dans les projections qu'il établit sur l'évolution du climat, le GIEC ne tient généralement compte que de l'influence sur le climat de l'augmentation des gaz à effet de serre imputable aux activités humaines et d'autres facteurs liés à l'homme. » L'évolution du climat au cours des prochaines décennies va dépendre de l'évolution des émissions de gaz à effet de serre et donc de l'évolution des sociétés humaines. Quatre grandes familles de scénarios ont été développées en combinant

différents changements démographiques, développements économiques et sociaux, et évolutions technologiques (IPCC Special Reports Emission Scenarios 2000).

- Famille A1 :
    - diminution des inégalités régionales ;
    - maximum démographique atteint vers le milieu du siècle, puis décroissance ;
    - croissance économique très rapide ;
    - introduction de nouvelles technologies.

Cette famille est divisée en catégories liées aux énergies utilisées majoritairement : fossiles (A1FI), non-fossiles (A1T) ou mixte (A1B).

- Famille A2 :
    - monde très hétérogène ;
    - croissance démographique continue ;
    - développement économique faible ;
    - progrès technologique lent.
- Famille B1 (proche de A1):
    - diminution des inégalités régionales ;
    - maximum démographique atteint vers le milieu du siècle ;
    - évolution vers une économie de communication et de services ;
    - introduction de technologies « propres », de haute efficacité énergétique.
- Famille B2 :
    - croissance démographique continue mais plus lente que pour A2 ;
    - niveau intermédiaire de développement économique ;
    - changements technologiques plus lents et plus variés que pour A1 et B1 ;
    - actions de protection environnementale et de progrès social à l'échelle régionale.

Les prédictions d'émissions de gaz à effet de serre sous ces différents scénarios sont utilisées pour alimenter différents modèles de circulation décrivant le fonctionnement de l'atmosphère terrestre. Les résultats obtenus font état d'une augmentation de la température comprise entre 1 et 6° d'ici la fin du siècle avec de très fortes incertitudes liées aussi bien aux scénarios d'émission de gaz à effet de serre qu'aux modèles utilisés (Figure 4).

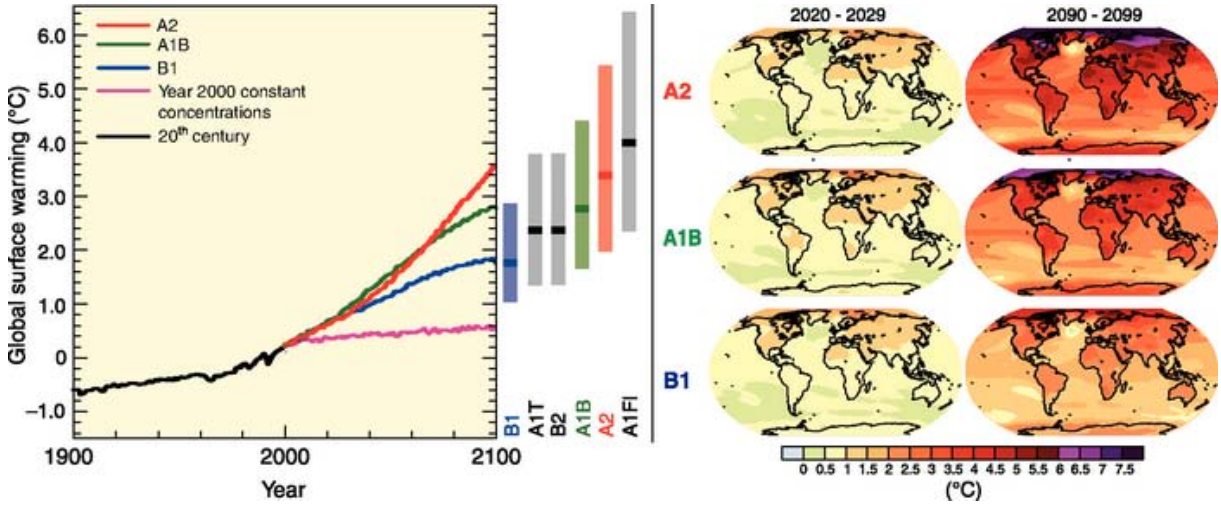


Figure 4 : (gauche) Réchauffement global de la température à la surface du globe par rapport à 1980-1999 (moyennes mondiales) pour trois scénarios de référence (B1 - bleu foncé, A1B - vert, A2 - rouge) en prolongement des simulations relatives au 20ème siècle. La courbe rose correspond à un maintien des concentrations atmosphériques de gaz à effet de serre aux niveaux de 2000. Les barres sur la droite précisent la fourchette probable d'augmentation de température pour la période 2090-2099 sous les six scénarios de référence. La zone foncée à l'intérieur des barres correspond à la valeur la plus probable de réchauffement. **(droite)** Prévision de changement de la température de surface pour le début et la fin du 21ème siècle par rapport à la période 1980-1999. Projections moyennes pour les scénarii A2 (haut), A1B (centre) et B1 (bas) pour les décennies 2020-2029 (gauche) et 2090-2099 (droite). Source : IPCC (2007).

4. Les impacts écologiques

Les impacts du réchauffement climatique sur la biodiversité sont nombreux et présents à toutes les échelles de l'organisme aux biomes (Figure 5). Le réchauffement climatique observé le siècle dernier, même s'il était de faible amplitude, a déjà eu des conséquences écologiques observables, en particulier sur la phénologie des espèces (Parmesan and Yohe 2003; Root et al. 2003). De nombreux évènements printaniers (floraison, ponte, migration) ont lieu plus précocement qu'avant, avec un décalage pouvant dépasser la semaine.

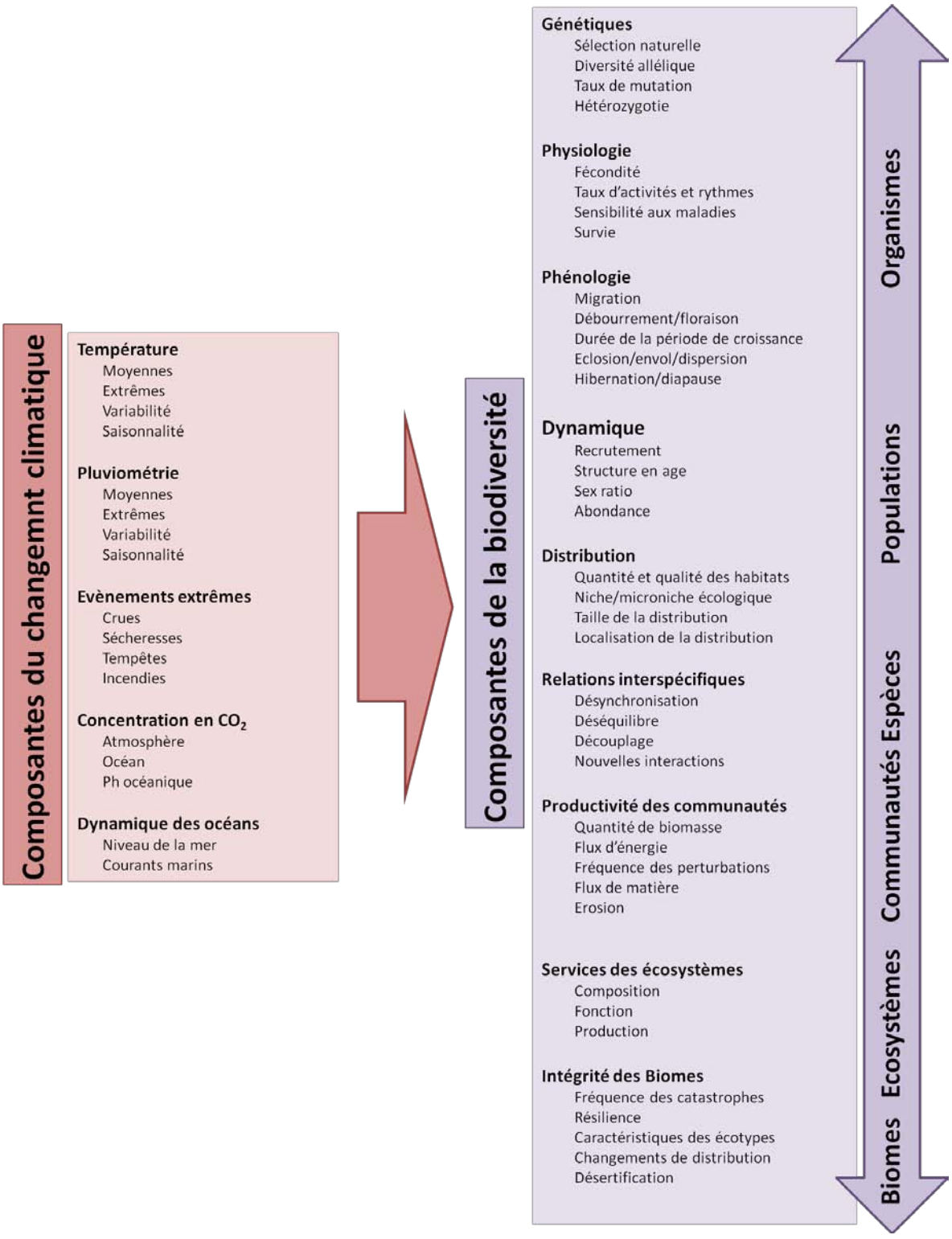


Figure 5 : Quelques-uns des aspects prévus du changement climatique et quelques exemples de leurs effets probables à différents niveaux de la biodiversité (d'après Bellard et al. 2012).

Une autre conséquence observable de l'élévation des températures est le déplacement des aires de distribution des espèces vers des latitudes (Parmesan et al. 1999) ou des altitudes plus élevées (Moritz et al. 2008; Rowe et al. 2010). Cette possibilité d'adaptation par migration des aires de distributions diffère suivant la localisation géographique. Certaines régions devraient être plus touchées par le changement climatique. C'est le cas de certaines régions très hétérogènes, à la limite entre plusieurs grandes régions biogéographiques (McInnes et al. 2009) mais aussi de régions isolées géographiquement, comme les îles ou les zones amont des cours d'eau, dans lesquelles le changement climatique conduit à une diminution voire une disparition des aires de distribution. De nombreuses espèces aussi bien animales que végétales risquent ainsi d'être mises en danger (e.g., Thuiller et al. 2005a; Barbet-Massin et al. 2009; Bond et al. 2011; Sauer et al. 2011; Crossman et al. 2012), puisque leur aire de distribution devrait se réduire très fortement.

Dans le cas d'un « simple » déplacement des aires de distribution, l'impact du changement climatique va dépendre de l'aptitude des espèces à se déplacer pour suivre les conditions environnementales adaptées à leurs besoins. La capacité de dispersion des espèces est donc un facteur déterminant dans l'impact potentiel du changement climatique. Si des études ne prenant pas en compte les contraintes de dispersion montrent que l'élévation des températures pourrait augmenter l'aire de distribution de certaines espèces (Araújo et al. 2006; Sharma and Jackson 2008; Barbet-Massin et al. 2009; Bond et al. 2011), l'intégration des contraintes de dispersion dans les modèles limite généralement beaucoup cet accroissement voire conduit à une prédiction de réduction de l'aire de distribution (Araújo et al. 2006; Buse and Griebeler 2011).

B. Les espèces invasives

1. Définitions

Le changement climatique n'est qu'une des perturbations qui affectent les écosystèmes. Depuis de nombreuses années, les espèces invasives ont été identifiées comme une des sources les plus importantes de modification de la planète par l'homme et une des causes principales de la disparition d'espèces, après la fragmentation et la disparition des habitats (Vitousek et al. 1996; Mack et al. 2000; Clavero and Garcia-Berthou 2005).

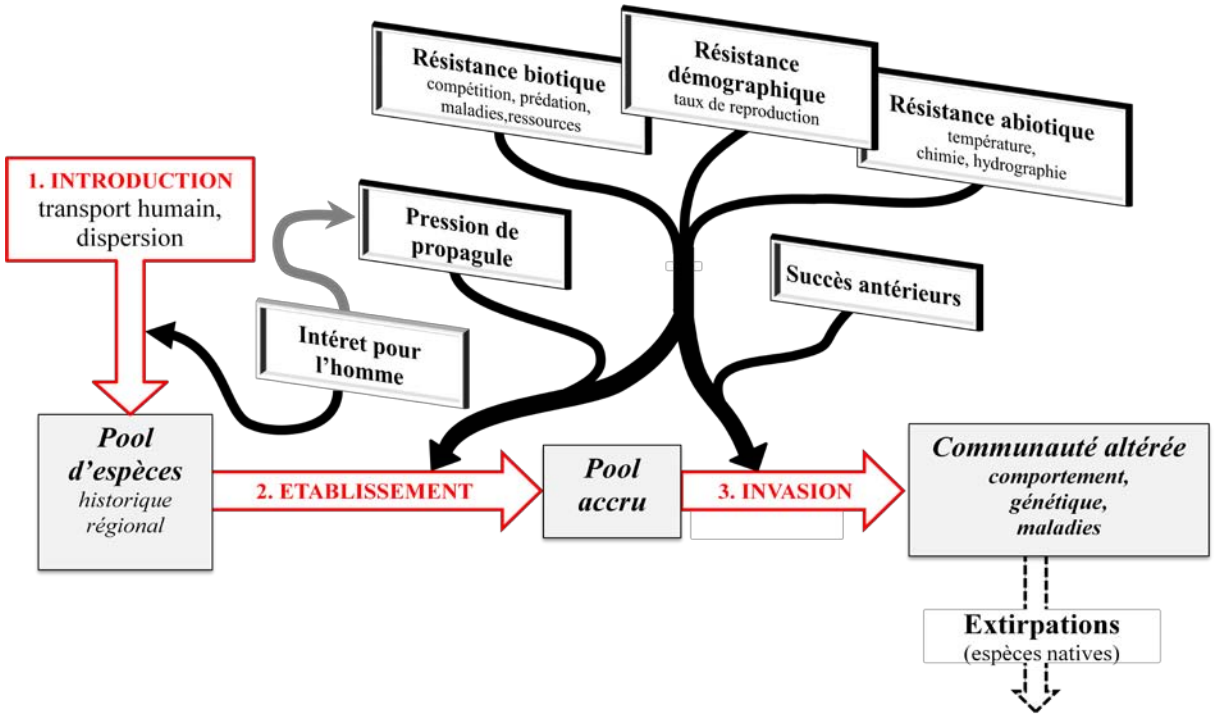


Figure 6 : Les différentes étapes du processus d'invasion, leurs effets sur la communauté et les facteurs contrôlant ces étapes.

On distingue en général trois phases dans le processus d'invasion (Williamson 1996; Richardson et al. 2000; Lockwood et al. 2007) (Figure 6). Dans un premier temps, l'homme permet à l'espèce d'arriver dans une zone géographique à l'extérieur de son aire native (c'est-à-dire hors de la zone géographique qu'elle peut occuper sans intervention anthropique). C'est l'**introduction**. Cette introduction peut être volontaire dans le cas d'espèces ornementales, d'intérêt économique ou dans le cadre de la lutte biologique. Elle peut aussi être accidentelle,

lors du transport d'autres espèces ou par la création de nouvelles voies de dispersion (canaux, ponts…). Durant la deuxième moitié du 20ième siècle, l'accroissement du commerce et des voyages, en augmentant la perméabilité des barrières géographiques qui limitent la dispersion des espèces, a intensifié le phénomène. L'activité humaine est d'ailleurs le principal facteur responsable des patrons mondiaux de richesse en espèces non-natives (Taylor and Irwin 2004; Leprieur et al. 2008; Pysek et al. 2010).

Une fois introduite, l'espèce peut se reproduire et former une population viable qui se maintient au cours du temps. C'est l'**établissement**. Le succès d'établissement dépend d'un grand nombre de paramètres biotiques et abiotiques. La pression de propagule (Cassey et al. 2004; Ruesink 2005; Jeschke and Strayer 2006; Yang et al. 2012), le « lien avec l'homme » (Marchetti et al. 2004b; Ruesink 2005; Jeschke and Strayer 2006) et l'adéquation des conditions environnementales entre la zone d'origine de l'espèce et la zone d'introduction (Blackburn and Duncan 2001; Moyle and Marchetti 2006; Bomford et al. 2009) apparaissent comme les facteurs essentiels du succès de cette étape. Les espèces généralistes (Cassey et al. 2004), à fécondité élevée, ou présentant des soins parentaux dans le cas d'espèces animales, semblent également s'établir plus facilement.

Une fois établie, l'espèce peut fonder de nouvelles populations viables, disperser largement et s'incorporer en grand nombre dans l'écosystème receveur. C'est l'**invasion**. Comme lors de l'établissement, les caractéristiques biotiques et abiotiques du milieu d'accueil contrôlent le succès du processus. Des conditions environnementales proches de celles de la zone native (Fausch et al. 2001; Moyle and Marchetti 2006), des perturbations anthropiques comme la pollution, l'urbanisation et l'agriculture augmentent les risques d'invasion (Marchetti et al. 2004a). L'effet de la richesse spécifique du milieu récepteur est quand à lui controversé. Si certaines études montrent qu'une forte richesse spécifique et la présence d'autres espèces invasives sont de bons indicateurs du succès de l'invasion (hypothèse d'acceptance

biologique, Moyle and Marchetti 2006), d'autres ont mis en évidence une tendance à la résistance biotique des espèces natives avec une aire de distribution des espèces exotiques plus réduite dans les régions à forte diversité (Guo et al. 2006). Certaines caractéristiques de l'espèce comme une forte tolérance physiologique (Moyle and Marchetti 2006; Segurado et al. 2011), un fort taux de reproduction (Ruesink 2005), un comportement agressif, explorateur et vorace (Conrad et al. 2011) et un succès invasif dans une autre aire géographique (Kolar and Lodge 2001; Marchetti et al. 2004b; Moyle and Marchetti 2006) sont également de bons indicateurs du succès probable de l'invasion. Cependant, l'identification des critères spécifiques d'invasiveté nécessite de prendre en compte l'aire de distribution potentielle de l'espèce et la date de début des introductions (Wilson et al. 2007). En effet, une espèce introduite depuis longtemps ou pouvant potentiellement s'établir sur de vastes étendues aura une distribution observée plus large que celle d'autres espèces aux caractéristiques similaires, mais introduites plus récemment ou à niche potentielle plus restreinte. Il est également important d'augmenter le nombre d'espèces étudiées, les espèces ayant fait l'objet d'études étant généralement les plus invasives et impactant le plus les écosystèmes receveurs (Pysek et al. 2008). Elles ne sont donc probablement pas représentatives de l'ensemble des espèces invasives.

Si la règle des 10% (Williamson 1996) prédit qu'à chaque étape (introduction, établissement, invasion) seul un dixième des espèces réussit à atteindre l'étape suivante, des études ont montré un taux de succès beaucoup plus élevé, en particulier chez les vertébrés (Jeschke and Strayer 2005). La disparition des barrières naturelles à la dispersion pourrait donc conduire à l'établissement d'espèces non natives dans de nombreux milieux, avec des impacts potentiels sur les écosystèmes receveurs d'autant plus importants que le nombre d'espèces invasives introduites risque d'être très élevé.

2. Les impacts

Les impacts des espèces invasives sont observables à toutes les échelles, du génome à l'écosystème et peuvent aller de modifications du comportement ou de l'habitat d'une ou de plusieurs espèces natives à la restructuration complète des réseaux trophiques. Les processus en jeu sont divers. Les espèces exotiques peuvent être le vecteur de maladies et de parasites jusque là absents du milieu (Mack et al. 2000; Clavero and Garcia-Berthou 2005). En s'hybridant avec les espèces natives, elles peuvent en diminuer la fitness, voire conduire à l'apparition d'hybrides beaucoup plus compétitifs que les espèces natives et qui finissent par les remplacer (Cucherousset and Olden 2011). En entrant en compétition avec les espèces natives, elles peuvent en diminuer la survie et en modifier le comportement (Blanchet et al. 2007). Ces changements de comportement peuvent conduire à une cascade de modifications au sein du réseau trophique. Par exemple, l'introduction de la truite fario *Salmo trutta* en Nouvelle Zélande a modifié le comportement des invertébrés brouteurs d'algues. Ces derniers ont réduit leur activité alimentaire pour éviter la prédation par la truite, a conduit à un accroissement des biomasses algales (Townsend 1996). Mais la truite a également eu des effets beaucoup plus directs sur les populations locales de poissons qui ont subi une forte pression de prédation, mettant en danger de nombreuses populations de galaxiidés endémiques de Nouvelle Zélande. La pression de prédation d'espèces exotiques a d'ailleurs parfois mené à des extinctions locales voire globales d'espèces natives (Townsend and Crowl 1991; Closs and Lake 1996; Clavero and Garcia-Berthou 2005). Elle a même, dans certains cas, eu des impacts non seulement sur l'écosystème receveur mais aussi sur les écosystèmes adjacents (Baxter et al. 2004). L'effet des espèces invasives sur la biodiversité et le fonctionnement des écosystèmes peut ainsi provoquer une perte importantes de services écologiques (Vitousek et al. 1997; Hooper et al. 2005) et donc avoir des conséquences économiques non négligeables (Pimentel et al. 2000; Hooper et al. 2005). Pour se faire une

idée de l'impact économique des espèces invasives, il suffit de voir que le coût estimé aux USA sur 10 ans du seul parasite invasif du chêne *Phytophthora ramorum* est de plus de 140 millions de dollars (Kovacs et al. 2011) et que le coût annuel des dommages et du contrôle des espèces invasives aux USA était évalué à plus de 138 milliards de dollars en 2005 (Pimentel et al. 2005).

3. La prévention

Une fois qu'une espèce est devenue invasive, son contrôle demande un engagement important et des efforts sur le long terme, de préférence à l'échelle de l'écosystème (Mack et al. 2000). De nombreuses pistes ont été envisagées pour améliorer le contrôle des espèces invasives : utilisation de l'effet Allee (croissance négative des populations d'effectif réduit), en intervenant lors des pics d'abondance de la population (Johnson et al. 2006), facilitation des adaptations comportementales locales des espèces natives (Schlaepfer et al. 2005) ou utilisation des modèles de distribution pour optimiser les campagnes de contrôle (Januchowski-Hartley et al. 2011). Mais dans tous les cas, l'éradication totale est extrêmement onéreuse et souvent impossible à mener à terme (Zambrano et al. 2006).

Il est donc capital de pouvoir lutter en intervenant de manière précoce, c'est-à-dire avant l'établissement, et si possible avant même la phase d'introduction (Moyle and Light 1996; Simberloff and Stiling 1996), soit en identifiant les espèces qui présentent un potentiel invasif élevé, soit en déterminant les zones géographiques à risque. Si certaines caractéristiques de l'espèce semblent faciliter l'invasion (soins parentaux, fort taux de reproduction, fortes capacités de dispersion, comportement explorateur), les études visant à établir des critères généraux permettant d'identifier les espèces à fort potentiel invasif se sont généralement révélées peu concluantes (Mack et al. 2000). Il est cependant possible de déterminer les régions susceptibles d'être envahies par des espèces déjà connues comme invasives dans

d'autres régions du globe. Les modèles de distribution sont un outil très utile non seulement pour déterminer les zones où les introductions sont à proscrire mais aussi pour cibler les régions à surveiller prioritairement dans le cas d'introductions accidentelles (Thuiller et al. 2005b).

C. Les modèles de niche

1. La définition de la niche

Hutchinson (1957) définit la niche comme un « ensemble de points dans un espace abstrait à n dimensions », chaque axe correspondant à une variable environnementale (biotique ou abiotique). Les points de la niche correspondent aux conditions environnementales qui permettent à l'espèce de maintenir des populations viables. Mais pour qu'une espèce soit présente en un lieu donné, il faut non seulement que les conditions environnementales lui conviennent (niche fondamentale), mais aussi que les interactions avec les espèces présentes (prédation, compétition, parasitisme…) permettent à l'espèce de se maintenir, et pour finir que l'espèce ait eu par le passé les moyens et le temps d'arriver dans ce lieu. Et même dans le cas où l'espèce est présente, cette présence ne peut être connue que si le lieu a été échantillonné. L'espèce n'est donc observée que dans une petite partie de sa niche (Figure 7).
Que ce soit dans le cadre de la protection des espèces, pour aider à la définition des zones de protection (Bombi et al. 2011) ou pour cibler les zones à échantillonner pour identifier des populations jusque là inconnues (Pearson et al. 2007; Williams et al. 2009), pour l'étude de l'impact du changement climatique ou pour la prévention des invasions, la zone que l'on cherche à modéliser est généralement soit la distribution potentielle, soit la niche toute entière. Mais les interactions entre espèces sont souvent difficiles à établir et leur prise en compte nécessite des données sur de nombreuses espèces et pas seulement l'espèce cible. De plus, sous l'effet du changement global, de nombreux assemblages vont subir des

changements, difficiles à prendre en compte sauf dans le cadre de modèles multi-spécifiques complexes. De nombreux modèles de distribution se fixent donc pour objectif de décrire la niche (abiotique) de l'espèce.

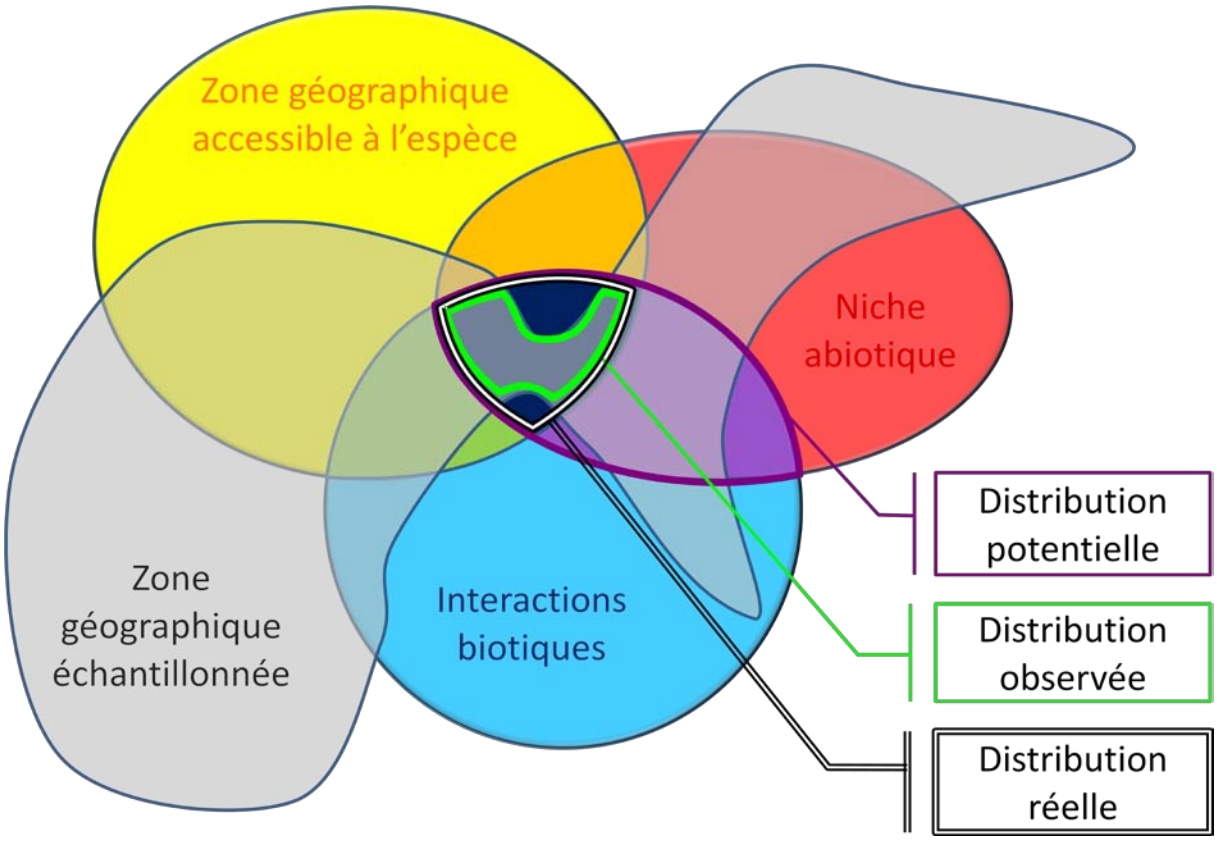


Figure 7 : Les différents types de distribution d'une espèce, modifié d'après (Soberón and Peterson 2005; Godsoe 2010).

2. Les modèles

Pour déterminer la niche d'une espèce, il existe deux grands types de modèles (Morin and Thuiller 2009). D'une part, les modèles mécanistiques (Kearney et al. 2008; Kearney and Porter 2009; Evans et al. 2012) sont basés sur les caractéristiques physiologiques et écologiques de l'espèce (Figure 8). Ils nécessitent une connaissance approfondie de l'espèce et ne sont donc applicables que dans le cas d'espèces dont la biologie est bien connue. Ils sont en général réservés à des espèces modèles en écologie ou de haute valeur économique.

D'autre part, les modèles corrélatifs sont basés sur l'établissement de relations statistiques entre des variables environnementales (climatiques, topographiques,…) et les distributions observées des espèces. Les moins exigeants ne nécessitent que des données de présence (presence-only models), et si les plus répandus sont ceux basés sur des données de présence/absence, on peut aussi construire des modèles corrélatifs basés sur des données d'abondance. Des études comparatives ont montré que les deux types d'approches (corrélative et mécanistiques) donnaient des résultats assez similaires (Kearney et al. 2010).

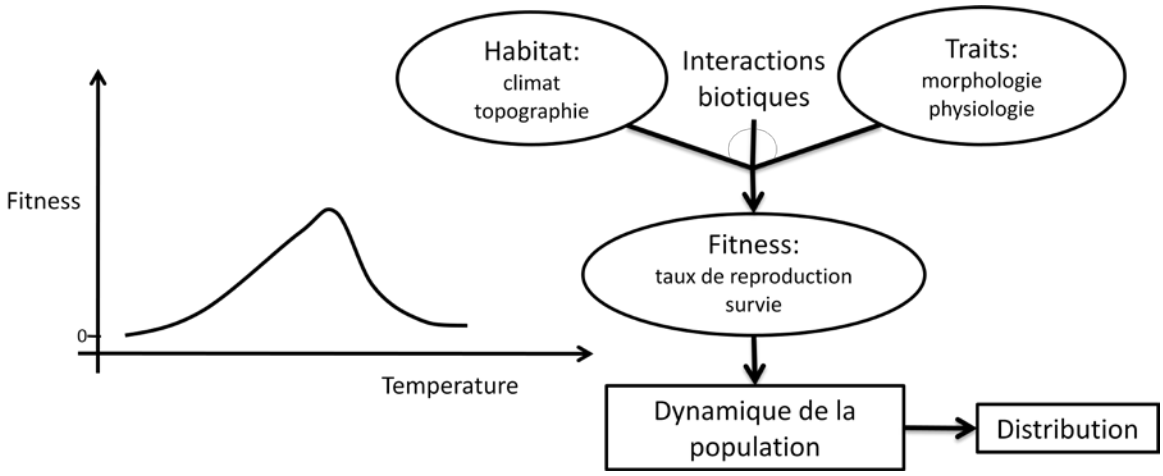


Figure 8 : Principe des modèles mécanistiques, basés sur la connaissance des processus physiologiques. Un modèle qui décrit explicitement les interactions entre les traits fonctionnels et les caractéristiques environnementales permet d'obtenir les paramètres de fitness (reproduction, survie) de l'espèce dans un environnement (biotique et abiotique) donné. On peut alors en déduire la dynamique de la population sous ces conditions environnementales et finalement estimer la distribution de l'espèce, i.e. les régions où l'espèce est susceptible de maintenir des populations viables.

Les modèles statistiques de présence/absence utilisés dans le cadre des modèles corrélatifs sont très variés : modèles linéaires généralisés (McCullagh and Nelder 1989), modèles additifs généralisés (Hastie and Tibshirani 1990; Hastie 2006), arbres de classification (Therneau and Atkinson 2007), Generalized Boosting regression Methods (Friedman 2001), forêts aléatoires (Breiman 2001; Liaw and Wiener 2002), analyse linéaire discriminante (Venables and Ripley 2002)…

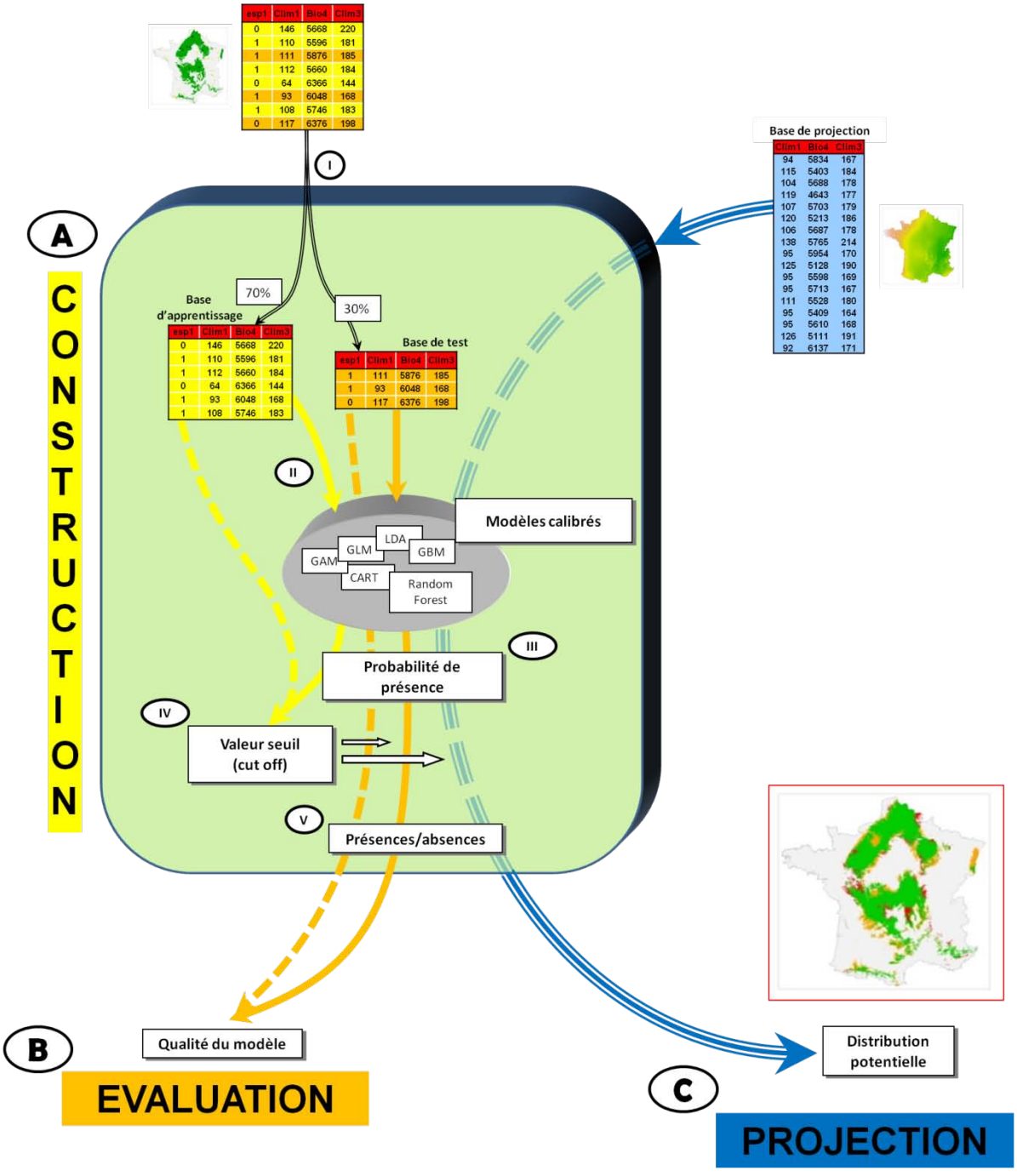


Figure 9 : Les principales étapes de la modélisation de niche par des modèles corrélatifs :
A) construction (jaune) ; B) évaluation (orange); C) projection (bleu). Les flèches pointillées correspondent à l'utilisation des données de présence/absence.

A. Construction des modèles

I. La base de données contenant les occurrences de l'espèce et les données environnementales est divisée aléatoirement en une base d'apprentissage destinée à la calibration des modèles et une base de test servant à l'évaluation de la qualité des modèles.

II. Les différents modèles statistiques utilisés sont calibrés en utilisant la base d'apprentissage.

III. Les modèles calibrés sont utilisés pour prédire les probabilités de présence de l'espèce pour tous les sites de la base d'apprentissage. Une probabilité de présence moyenne est ensuite calculée.

IV. Les occurrences et les probabilités de présence de la base d'apprentissage sont utilisées pour déterminer la valeur seuil en fonction du critère choisi.

V. Les probabilités de présences sont converties en données de présence/absence en utilisant la valeur seuil déterminée en IV.

B. Evaluation

Les données de la base de test et les modèles construits en A sont utilisés pour prédire une probabilité de présence sur l'ensemble des sites de test. Cette probabilité de présence est convertie en 0/1 en utilisant la valeur seuil déterminée en A. Les occurrences de la base de test et les prédictions (probabilité de présence et présence/absence) servent à évaluer la qualité des modèles en utilisant des mesures indépendantes (AUC) et dépendantes (Kappa, TSS) de la valeur seuil.

C. Projection

Les données de la base de projection et les modèles construits en A sont utilisés pour prédire une probabilité de présence sur l'ensemble des sites de la zone d'étude. Cette probabilité de présence est convertie en 0/1 en utilisant la valeur seuil déterminée en A. Les prédictions de présence/absence permettent d'établir la carte de la distribution potentielle de l'espèce.

Le processus complet peut être répété plusieurs fois pour prendre en compte la variabilité des prédictions due au choix de la base d'apprentissage, en particulier pour les espèces peu fréquentes, très fréquentes ou lorsque le nombre total d'observations dans la base de données est faible, le choix de la base d'apprentissage affectant alors fortement les modèles.

Les résultats de ces différentes techniques varient beaucoup aussi bien en termes de performance que d'aire prédite et malgré un très grand nombre d'études, aucun consensus n'a émergé sur la ou les méthodes fournissant les meilleures prédictions. Le choix de la méthode statistique est même la première source de variabilité des prédictions sous l'effet du changement climatique, bien avant le scénario d'émission de gaz à effet de serre ou le modèle de circulation choisis (Buisson et al. 2010). Il semble donc préférable d'utiliser des méthodes d'ensemble (ensemble modelling, EM, Araújo et al. 2005), qui utilisent simultanément plusieurs méthodes et améliorent grandement la performance des modèles prédictifs de distributions (Species Distribution Models ou SDMs) (Araújo and New 2007; Marmion et al. 2009; Stohlgren et al. 2010; Grenouillet et al. 2011).

En plus de prédire les zones potentielles d'établissement, les modèles de distribution présentent d'autres possibilités d'application dans le cadre des espèces invasives. Ils permettent d'optimiser les campagnes de lutte en estimant les zones où l'élimination a été la plus efficace et en ciblant les zones préférentielles d'intervention (Januchowski-Hartley et al. 2011). Ils sont également utilisés dans des modèles hybrides qui couplent des modèles mécanistiques et des modèles corrélatifs. Les modèles hybrides les plus utilisés utilisent les prédictions des modèles corrélatifs (probabilité de présence) pour contraindre les paramètres des modèles mécanistiques (survie, taux de dispersion,…). Leur usage semble très prometteur pour la prédiction de la distribution potentielle des espèces invasives, la dynamique de leurs populations et les conséquences du processus d'invasion (Gallien et al. 2010). Malheureusement, comme les modèles mécanistiques qu'ils intègrent, ils nécessitent une connaissance approfondie des processus en jeu et ne peuvent donc s'appliquer qu'à un nombre réduit d'espèces bien étudiées.

Malgré leur large champ d'application et leur efficacité avérée, les modèles corrélatifs, qu'ils soient basés sur des données de présence ou sur des données de présence/absence, rencontrent

un certain nombre de problèmes liés à leur construction. Certains sont directement liés aux processus de modélisation, d'autres à la nature des données utilisées pour construire les modèles.

3. Les problèmes des modèles corrélatifs liés à la modélisation

a) Le choix des variables environnementales

Le choix des variables environnementales est une étape délicate du processus de modélisation, en particulier dans le cadre des prédictions sous l'effet du changement climatique, puisque les variables sélectionnées sont une source majeure de variabilité des prédictions (Synes and Osborne 2011). Si des considérations écologiques permettent d'éliminer les variables peu pertinentes pour l'espèce considérée, le choix final reste délicat. Les modèles construits avec trop peu ou trop de variables sont souvent de moins bonne qualité (Warren and Seifert 2011). Un trop petit nombre de variables ne permet pas de capturer l'ensemble des contraintes définissant la niche. A l'opposé, l'augmentation du nombre de variables diminue la précision de l'estimation des paramètres en augmentant les corrélations entre variables et fait aussi courir le risque de sur-apprentissage, ce qui limite les qualités prédictives des modèles. Les modèles prenant en compte un grand nombre de variables doivent être réservés à la détermination de la niche observée (Jiménez-Valverde et al. 2008). Les méthodes statistiques classiques de sélection de variables (backward stepwise variable selection, Akaike Information Criteria) ne fournissent qu'une aide relative, car dans le cadre de l' « ensemble modelling » les variables sélectionnées par les algorithmes de modélisations peuvent dépendre du modèle statistique choisi (Syphard and Franklin 2009). De plus la sélection des variables peut être influencée par la présence de fausses absences (l'espèce n'a pas été détectée alors qu'elle est présente) dans la base d'apprentissage (Cianfrani et al. 2010) ou par l'échelle (Millar and Blouin-Demers 2012). Le choix des

variables environnementales doit donc faire intervenir à la fois des considérations écologiques (sélection en fonction des besoins de l'espèce) et des contraintes statistiques (variables peu corrélées) tout en limitant le nombre de variables sélectionnées.

b) Le choix des valeurs seuil

Les modèles de présence/absence prédisent pour chaque site une probabilité de présence de l'espèce (un niveau d'habitabilité). Si ces résultats sont facilement utilisables pour comparer l'habitabilité des différents sites pour une même espèce, ils sont difficilement utilisables pour les comparaisons entre espèces car la distribution de probabilités est fortement corrélée à la prévalence de l'espèce. Ils ne permettent également pas de mesurer des changements de caractéristiques de l'aire de distribution (surface, gamme altitudinale) entre différentes périodes. Les résultats sont donc souvent convertis en données de type présence/absence par l'utilisation d'un seuil ce qui permet une uniformisation des résultats (Jiménez-Valverde and Lobo 2006). Il existe différentes méthodes pour déterminer ce seuil (Liu et al. 2005). Certaines, comme l'utilisation d'un seuil arbitraire de 0.5, ont été abandonnés (Liu et al. 2005; Jiménez-Valverde and Lobo 2007; Freeman and Moisen 2008). De nombreuses études favorisent la maximisation de la somme de la sensitivité et de la spécificité (ou de façon équivalente la TSS) (Liu et al. 2005), même si cette méthode tend à surestimer l'aire prédite (Manel et al. 2001), d'autres études conseillent la prévalence (Liu et al. 2005; Freeman and Moisen 2008) ou la maximisation du Kappa (Freeman and Moisen 2008), mais aucun consensus définitif n'émerge. Cependant le choix de la valeur seuil est le second facteur le plus important pour expliquer la variabilité des projections de distribution sous l'effet du changement climatique après la méthode statistique utilisée et bien avant le scénario de changement climatique (Nenzen and Araujo 2011).

c) La mesure de la qualité des modèles

Une fois les modèles construits, il est important de pouvoir évaluer la fiabilité des résultats obtenus. La mesure indépendante du seuil la plus utilisée (évaluée sur les probabilités de présence) est l'AUC (Hanley and McNeil 1982) : aire sous la courbe ROC (Receiver Operating Characteristics). Mais cette mesure fait l'objet de certaines critiques (Lobo et al. 2008; Cianfrani et al. 2010), en particulier parce qu'elle prend en compte des zones de la courbe qui ont peu de sens écologique et qu'elle dépend fortement de l'étendue de la zone sur laquelle les modèles sont construits. Une mesure de l'aire partielle sous la courbe a également été proposée (Peterson et al. 2008) pour ne tenir compte que de la zone de la courbe écologiquement pertinente, mais elle reste à ce jour peu utilisée.

Le consensus est encore moins de mise avec les mesures sur les prédictions binaires (présence/absence). Les mesures les plus utilisées sont le pourcentage de sites bien prédits, la sensitivité, la spécificité, la TSS et le Kappa. Tous ces indices mesurent la qualité des modèles de façon différente, en particulier du fait de leur dépendance à la prévalence (Mouton et al. 2010). Le Kappa par exemple, d'abord conseillé (Manel et al. 2001), a été ensuite fortement décrié du fait de valeurs parfois très élevées malgré une sur ou sous-estimation très importante de l'aire et une dépendance très nette à la prévalence (Allouche et al. 2006; Santika 2011). Il a finalement été réhabilité puisque cette dépendance à la prévalence traduit simplement sa façon de prendre en compte l'effet du hasard (Santika 2011). Les différents indices mesurant des aspects différents de la qualité des modèles (Mouton et al. 2010), on peut choisir d'en utiliser plusieurs, ou de cibler ceux qui sont le plus adaptés aux objectifs de l'étude.

Le fait que certains indices servent à la fois dans la sélection de la valeur seuil et dans l'évaluation de la qualité des modèles ne facilite pas le choix des indices ni l'interprétation des résultats. Par construction, un indice de mesure de qualité aura tendance à avoir des valeurs plus élevées pour les modèles obtenus en maximisant cet indice pour choisir la valeur

seuil, même en évaluant la qualité sur une base de test relativement indépendante de la base utilisée pour calibrer les modèles.

Cette indépendance n'est d'ailleurs presque jamais vérifiée. Les bases d'apprentissage et de test sont généralement obtenues en séparant aléatoirement les sites d'échantillonnages en deux sous-ensembles. Les sites correspondants sont donc issus de la même zone géographique ce qui pose des problèmes de corrélation spatiale. Ce problème peut, en théorie, être résolu en construisant les modèles à partir de données issues d'une zone géographique et en les testant sur une zone géographique différente (mesure de transférabilité des modèles). Cette méthode est cependant déconseillée par certains (Newbold et al. 2010) et n'est envisageable que si les deux parties de la niche couvrent chacune l'ensemble des conditions environnementales. Dans le cas contraire, l'utilisation d'une partie seulement de la niche dans la phase de construction des modèles conduit à une perte d'information écologique et donc à l'obtention de modèles de qualité réduite (Randin et al. 2006). On rejoint là les problèmes liés à la qualité des données utilisées. Si au contraire, c'est la base de test qui ne couvre qu'une partie seulement de la niche, la qualité des modèles dans les conditions non représentées ne sera pas évaluée.

4. Les problèmes des modèles corrélatifs liés aux données

a) L'extrapolation

Les modèles corrélatifs sont construits sur une gamme de valeurs des variables environnementales qui correspond aux conditions rencontrées dans la zone échantillonnée et qui ne couvre donc pas l'ensemble des conditions environnementales possibles. La projection des modèles sur des conditions environnementales différentes, soit pour déterminer la distribution de l'espèce dans une autre zone géographique, soit à cause du changement global (Elith et al. 2010), pose un problème d'extrapolation (Webber et al. 2011) et peut conduire à des résultats erronés (Thuiller et al. 2004) .

b) Les biais d'échantillonnage

Les données d'occurrences sont souvent liées à un petit nombre de campagnes d'échantillonnages. Même lorsque ces campagnes sont basées sur des protocoles rigoureux, les zones difficiles d'accès sont généralement sous représentées. Or ces biais d'échantillonnage peuvent fortement impacter les prédictions des modèles. Cela a été montré en utilisant des données historiques et en comparant les modèles obtenus à partir des données de différentes périodes (Lobo et al. 2007; Hortal et al. 2008). Même quand les différents modèles donnent une image précise de la niche actuelle, les prédictions pour le futur peuvent présenter une grande variabilité (Stankowski and Parker 2011). Ces biais semblent néanmoins pouvoir être réduits en optimisant la complexité des modèles (Anderson and Gonzalez Jr 2011).

c) Les fausses absences

Les deux problèmes précédents sont liés aux régions échantillonnées, et peuvent être contrôlés, au moins partiellement. La nature écologique des données introduit un biais supplémentaire qui s'avère beaucoup plus difficile à éviter. En effet, les modèles corrélatifs sont basés sur l'hypothèse que l'espèce étudiée est à l'équilibre, autrement dit qu'elle occupe l'ensemble des conditions qui lui sont favorables. Mais cette hypothèse n'est pour ainsi dire jamais vérifiée. Si la présence d'une espèce est factuelle, avec des risques d'erreurs limités, par exemple suite à une mauvaise identification de l'espèce ou de par la présence occasionnelle de l'espèce en dehors de sa niche (déplacement entre deux patchs, crue ayant entrainé les organismes…), son absence est beaucoup plus incertaine et peut avoir des origines multiples (Lobo et al. 2010). L'espèce peut être réellement absente car les conditions environnementales ne lui conviennent pas ; ces absences sont alors informatives d'un point de vue écologique, et utiles pour déterminer la niche de l'espèce. Elle peut aussi avoir été

extirpée (lors des changements climatiques passés, par action anthropique,…) et ne pas avoir eu la possibilité de s'établir à nouveau. L'absence de l'espèce peut aussi être due à des interactions biotiques qui empêchent son établissement ou à des contraintes de dispersion (l'espèce n'a jamais eu l'occasion de s'établir dans ce milieu car elle n'y est jamais arrivée). Ces informations sont alors utiles si on cherche à étudier ou à modéliser la distribution réelle de l'espèce mais gênantes pour l'étude de la distribution potentielle de l'espèce, en particulier dans le cadre de l'étude des espèces invasives. Enfin, l'espèce peut ne pas avoir été détectée alors qu'elle est présente. Ces dernières absences sont les plus nuisibles à la détermination de la niche et peuvent représenter une part importante des données au sein de l'aire de distribution dans le cas d'espèces difficilement détectables, comme par exemple les grands carnivores en zone forestière (Cubaynes et al. 2010) ou certaines espèces de poissons difficilement détectables avec les méthodes d'échantillonnage classiquement utilisées.

5. Le changement de niche

L'utilisation des modèles corrélatifs pour prédire la distribution des espèces invasives rencontre une dernière difficulté qui lui est propre. En effet, si la niche fonctionnelle semble conservée au cours du processus d'invasion, la niche climatique ne l'est pas nécessairement lors de l'établissement de nouvelles populations par une espèce invasive (Larson et al. 2010). Ce changement de niche peut avoir plusieurs origines. L'espèce peut être adaptée à des conditions environnementales plus variées que celles de son aire native mais ne pas les occuper à cause d'interactions avec d'autres espèces ou simplement parce que ces conditions n'existent pas dans la zone accessible à l'espèce. Elle peut aussi s'adapter aux nouvelles conditions environnementales par plasticité ou mutation.

Si on n'a pas observé de façon directe des changements notables dans les préférences environnementales des espèces invasives, on dispose de preuves d'un potentiel d'adaptation rapide et de différentiation des populations aussi bien en laboratoire (Sexton et al. 2002) que sur le terrain (Maron et al. 2004). On dispose également d'au moins un exemple d'adaptation morphologique lors du processus d'invasion. Les populations de crapaud buffle (*Buffo marinus*) trouvés sur le front d'invasion australien présentent des jambes plus longues et une vitesse de dispersion plus élevée que celles établies depuis longtemps.

Des preuves indirectes de changement de niche existent, via l'utilisation des modèles de niche, en comparant les aires prédites par les modèles calibrés soit sur l'aire native, soit sur l'aire exotique (Fitzpatrick et al. 2007; Urban et al. 2007; Broennimann and Guisan 2008; Beaumont et al. 2009; Medley 2010; Villemant et al. 2011; Hill et al. 2012), ou en comparant les aires prédites à partir de données correspondant à différents stades du processus d'invasion (Václavík and Meentemeyer 2011; Hill et al. 2012). Cependant, les résultats obtenus semblent très sensibles au choix des variables sélectionnées lors de la construction des modèles et les changements de niche observés pourraient n'être que des artefacts liés à l'utilisation d'un trop grand nombre de variables par rapport au nombre d'occurrences disponibles (Peterson 2011) ou de différences d'amplitude des variables climatiques entre les deux régions (Hill et al. 2012). D'autre part, des couplages de modèles de distribution avec des études génétiques sur le lézard des murailles *Podarcis muralis* montrent que le changement de niche observé serait avant tout dû à une différence dans les niches réalisées avec conservation de la niche fondamentale (Schulte et al. 2011). Aucun consensus ne semble donc émerger. De nouvelles études sont souhaitables afin d'identifier les espèces qui n'ont pas une niche stable et augmenter ainsi la fiabilité des prédictions des modèles.

D. Le milieu aquatique d'eau douce

Alors que les milieux d'eau douce ne représentent que 1% de la surface du globe, contre 70% pour le milieu marin, ils contiennent près de la moitié des espèces de poissons (Leveque *et al.*, 2008) et sont particulièrement exposés aux changements globaux. S'ils sont déjà soumis à de nombreuses perturbations d'origine anthropique (fragmentation par de nombreux barrages (Nilsson et al. 2005), pollution agricole ou industrielle (Richter et al. 1997) ils sont aussi parmi les plus susceptibles de souffrir du changement climatique qui risque de provoquer la raréfaction voire la disparition de nombreuses espèces (Xenopoulos and Lodge 2006). Ces risques d'extinction sont liés à la nature insulaire des cours d'eau qui réduit très fortement la dispersion inter-bassins pour les espèces strictement inféodées à l'eau douce et sans stade terrestre (soit plus de 80% des poissons d'eau douce). De plus, au sein de chaque bassin, la structure dendritique des cours d'eau limite également la dispersion intra-bassin. Les espèces des zones amont sont alors les plus touchées car elles ne tolèrent souvent qu'une faible gamme thermique et ne peuvent pas fuir vers des zones plus froides. Au contraire, les espèces colonisant les zones aval sont généralement plus tolérantes d'un point de vue thermique et voient souvent leur aire s'élargir (Chu et al. 2005; Buisson et al. 2008; Rahel and Olden 2008; Heino et al. 2009; Domisch et al. 2011).

Les écosystèmes d'eau douce sont également susceptibles d'être impactés par l'introduction d'espèces non natives (Leprieur et al. 2008). Si les impacts potentiels des espèces invasives sur les milieux aquatiques ne sont pas toujours bien évalués (Crawford and Muir 2008; Leprieur et al. 2009a), de nombreuses études ont mis en évidence des effets négatifs (Richter et al. 1997; Wilcove et al. 1998; Canonico et al. 2005; Moyle and Marchetti 2006), en particulier des espèces piscivores (Mitchell and Knouft 2009). Ces effets, autant directs

qu'indirects, peuvent affecter significativement de nombreux taxa, à tous les niveaux d'organisation, du génome à l'écosystème (Cucherousset and Olden 2011) (Figure 10).

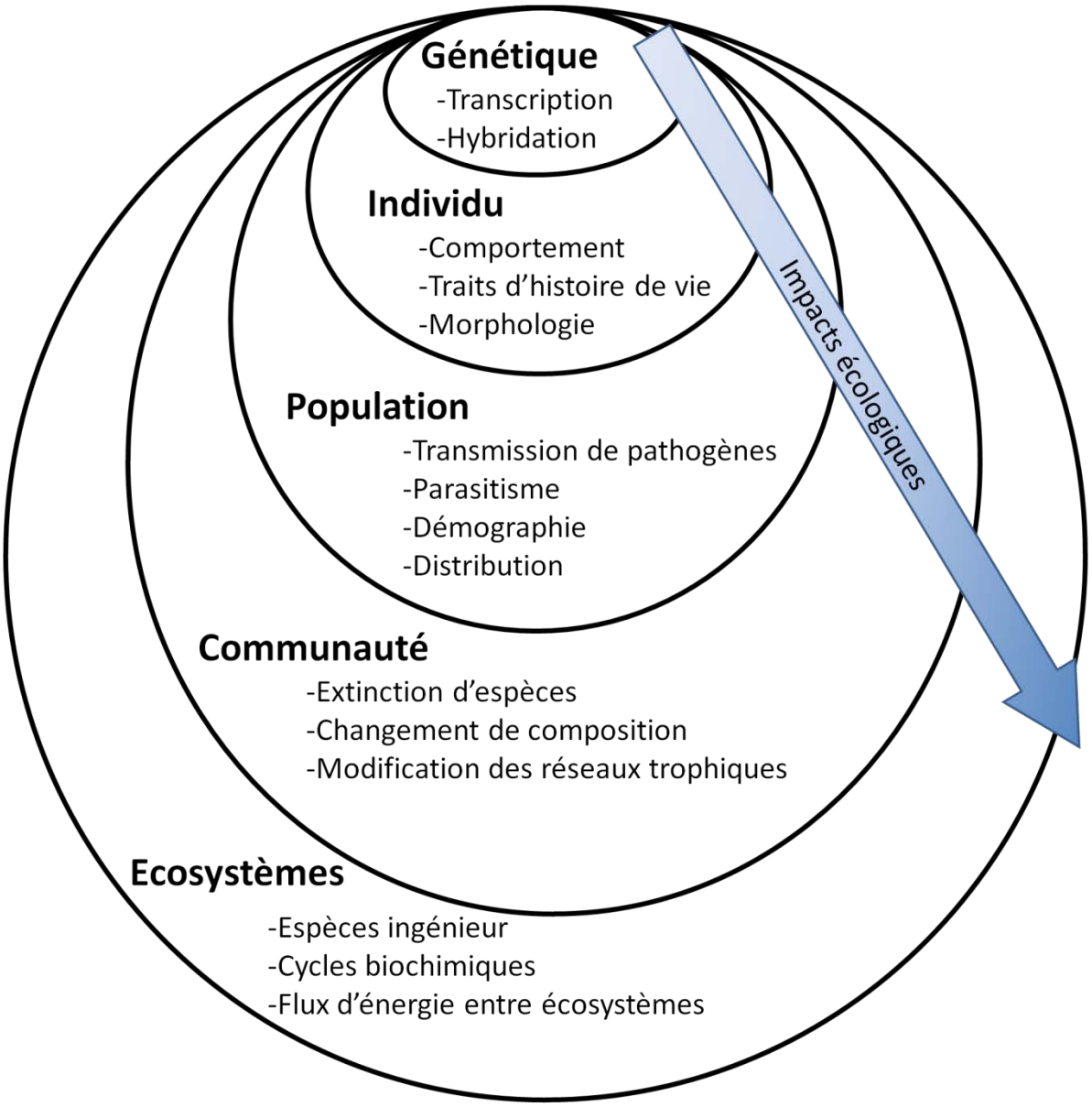


Figure 10 : Représentation schématique des impacts écologiques des espèces non natives de poissons d'eau douce à cinq niveaux d'organisation biologique. La flèche indique que les impacts des espèces non natives de poissons ne sont souvent pas limités à un niveau mais ont des répercussions à travers plusieurs niveaux d'organisation biologique (d'après Cucherousset and Olden 2011).

Même si on a noté quelques introductions accidentelles de poissons dans le milieu naturel via les eaux de ballast (Wonham et al. 2000) et si le creusement de canaux a créé des connections entre bassins jusque là séparés (Rahel 2007), les introductions de poissons sont généralement

dues à des libérations accidentelles après une importation volontaire sur le territoire dans le cadre de l'élevage piscicole (Crawford and Muir 2008), ainsi qu'à des introductions délibérées (espèces lâchées pour la pêche sportive). Les espèces d'intérêt faisant l'objet d'introductions multiples, et ce dans des régions variées, on observe un succès très élevé d'établissement chez les poissons: plus de la moitié des espèces introduites s'établissement, loin de la règle des 10% (Garcia-Berthou et al. 2005).

Les actions de lutte contre les invasions sont d'autant plus urgentes que :

- les introductions ont souvent lieu dans des hotspots de diversité ou des zones possédant des espèces en danger (Leprieur et al. 2008) ;
- on observe une accélération du nombre des espèces introduites ainsi que de la fréquence des introductions (Vander Zanden 2005; Clavero and Garcia-Berthou 2006) ;
- le réchauffement climatique devrait favoriser les espèces invasives, en particulier en motivant la construction de nouveaux barrages qui servent souvent de réservoirs à ces espèces mais aussi en favorisant leur dispersion suite à des crues de plus forte amplitude (Rahel and Olden 2008). L'établissement sera quant à lui localement facilité dans les milieux où le changement climatique aura provoqué le déclin de la faune native ;
- certaines espèces présentent une capacité d'invasion dramatiquement élevée, la Gambusie (*Gambusia sp.*) étant par exemple capable d'établir des populations viables à partir d'une seule femelle gravide (Deacon et al. 2011).

La nature "volontaire" des introductions de poissons devrait permettre de limiter les risques de nouvelles invasions, par le contrôle du commerce et des espèces de poissons autorisées à l'élevage. La prévention passe donc une fois encore par l'identification des espèces risquant de

s'acclimater dans la zone d'intérêt, dans les conditions actuelles ou sous l'effet du changement climatique, et donc par la détermination des distributions potentielles des espèces concernées.

E. Plan du manuscrit

Cette thèse est basée sur quatre manuscrits et un travail effectué dans le cadre du programme INVAQUA (INVAsion des milieux AQUAtiques) de l'Office National des Eaux et des Milieux Aquatiques (ONEMA) :

(M1) Geometry drives the grain size effects in species distribution models

Echelle : France
Données :

- 3 espèces virtuelles
- Données worldclim

Grain : 30"x30"→32'x32'
Résultats :

- Le grain utilisé a peu d'influence sur la qualité des modèles de niche.
- Le grain n'influe sur l'aire de la distribution prédite que par l'accroissement géométrique de l'aire de la distribution observée.
-

(M2) Identifying climatic niche shifts using coarse-grained occurrence data: a test with non-native freshwater fish

Echelle : Monde
Données :

- 18 espèces de poissons (base SPRICH)
- Données worldclim

Grain : 30"x30"
Résultats :

- Les données à large échelle sont utilisables pour identifier le changement de niche.
- Le changement de niche semble être un phénomène répandu chez les poissons d'eau douce.

Programme INVAQUA (étude des risques d'établissement de 6 espèces de poissons destinées à l'aquaculture et à la pêche sportive)

Echelle : France
Données :

- 6 espèces de poissons (Faunafri, NAWQA, DeVaney et al. 2009)
- Données worldclim, CGCM, HadCM

Grain : 30"x30"
Résultats :

- Les modèles d'ensemble prédisent correctement les sites d'établissement actuel de *Micropterus salmoides* en France métropolitaine.
- Sous l'effet du changement climatique, la plupart des espèces étudiées seront susceptibles de s'établir en France, au moins sur une partie du territoire.

(M3) Dealing with noisy absences to optimize species distribution models: an iterative ensemble modelling approach

Echelle : France
Données :

- 3 espèces virtuelles
- Données worldclim

Grain : 30"x30"
Résultats : La méthode itérative permet d'améliorer la qualité des modèles de distribution dans le cas de nombreuses absences non environnementales.

(M4) The iterative ensemble modelling approach increases the accuracy of fish distribution models

Echelle : France
Données :

- 31 espèces de poissons (données ONEMA)
- CRU CL 2.0

Grain : 10"x10"
Résultats : La méthode itérative permet d'améliorer la qualité des modèles de distribution des espèces difficiles à détecter.

L'ensemble des travaux effectués dans le cadre de ma thèse concerne la prédiction des risques d'invasion, en particulier sous l'effet du réchauffement climatique. Ces travaux portent à la fois sur des aspects méthodologiques (influence du grain sur la qualité des modèles (M1), amélioration de la méthode d'ensemble (M3, M4)) et sur des applications pratiques dans le cadre de la lutte contre les espèces invasives de poissons (identification des changements de niche à l'échelle du bassin (M2) ; prédiction des risques d'établissement d'espèces d'aquaculture en France (projet INVAQUA)).

La synthèse de mes travaux se décompose en trois parties. La première partie concerne l'utilisation de différents grains dans la prédiction des risques d'invasion et porte sur deux aspects différents de cette problématique. Dans un premier temps (M1), je testerai l'effet du grain sur les prédictions des modèles de distribution du point de vue de la qualité des modèles (mesurée en utilisant quelques indices classiques) ainsi que de la surface de la distribution prédite. Je montrerai que le choix du grain le plus fin n'est pas toujours le plus pertinent pour la construction des modèles corrélatifs et que la perte de qualité des modèles n'est nettement observable que pour les grains les plus grossiers. Dans un deuxième temps (M2), je regarderai si les données à très large grain (bassin versant), peuvent permettre d'identifier des changements de niche d'espèces invasives.

Dans une deuxième partie, j'utiliserai des méthodes d'ensemble « classiques » pour modéliser les distributions potentielles de 6 espèces de poissons dans les conditions climatiques actuelles et sous l'effet du changement climatique. Cette étude a été réalisée en collaboration avec le Muséum d'Histoire Naturelle de Paris suite à une demande de l'ONEMA dans le cadre du projet INVAQUA portant sur les risques d'établissement sur le territoire français de quelques espèces de poissons déjà introduites ou risquant d'être introduites pour l'aquaculture.

Dans une troisième partie, je présenterai une nouvelle méthode que j'ai développée pour améliorer les prédictions des modèles de distribution dans le cas où la base de données contient de nombreuses absences non environnementales. Cette méthode, inspirée de méthodes de simulations utilisées en physique, est basée sur des itérations successives, les sorties des modèles étant utilisées pour « corriger » les données d'occurrences utilisées pour calibrer les modèles à l'itération suivante. La méthode a été testée dans un premier temps sur des espèces virtuelles afin de contrôler précisément le nombre et la localisation des absences non environnementales. La méthode a ensuite été évaluée en utilisant des espèces réelles dont la niche est généralement plus difficile à modéliser. Dans ce dernier cas, les sites de la base de test ont été sélectionnés pour contenir un nombre réduit d'absences non environnementales et la quantité d'absences non environnementales contenue dans la base d'apprentissage a été estimée en mesurant la détectabilité des espèces.

Je terminerai par un chapitre de conclusions générales et de perspectives émergeant des principaux résultats obtenus.

Partie I :
De l'utilisation des données à grain large

Le grain des données utilisées pour construire les modèles de distribution est très variable, de 25mx25m à 50kmx50km dans le cas de la modélisation des distributions de plantes (Franklin 2009). Cette variété dans le grain utilisé est entre autres liée aux sources des occurrences qui peuvent être très variées (atlas, muséums, campagnes exhaustives d'échantillonnage…), en particulier dans le cas des espèces invasives pour lesquelles les régions dans lesquelles l'espèce est présente sont très diverses. Afin de construire les modèles à partir de données homogènes, on peut choisir de n'utiliser qu'une partie des données, mais cela peut diminuer la qualité des modèles en augmentant les biais d'échantillonnage (Feeley and Silman 2011). L'utilisation du maximum de données disponibles semble donc préférable. Le manque de précision dans la localisation géographique de certains sites (par exemple dans le cas de données de muséum pour lesquelles on ne connait que la commune ou la région de prélèvement) oblige alors à travailler à un grain large. Il est donc important de connaître l'impact du grain sur les prédictions des modèles de distribution mais aussi de savoir si l'utilisation des données à grain trop grossier pour être utilisées dans les modèles de distribution reste pertinente dans le cadre de la lutte contre les espèces invasives en permettant l'identification des changements de niche.

A. L'influence du grain sur la qualité des modèles (M1)

Le grain de l'étude n'est pas seulement contraint par les données d'occurrence des espèces mais aussi par les données climatologiques et topographiques. En particulier, les scénarii de changement climatique ne sont parfois disponibles qu'à un grain plus large que celui des données biologiques. Suivant les cas, on peut envisager, soit de construire les modèles à un grain donné et de les projeter à un autre (upscalling si le grain de projection est plus large, downscalling s'il est plus fin), soit de transformer les données pour utiliser dans tous les cas le grain le plus grossier.

Kriticos et Leriche (2010) ont mis en évidence une influence nette du changement du grain utilisé entre la construction des modèles et leur projection, l'upscalling ayant tendance à réduire la niche prédite, le downscalling à l'augmenter. Les études portant sur l'effet du grain à la fois pour la construction et la projection des modèles donnent, quant à elles, des résultats contradictoires. Alors que Guisan *et al.* (2007) ont montré que le grain avait une influence réduite sur la qualité des modèles, Seo *et al.* (2009) et Hu & Jiang (2010) ont observé un accroissement très important de l'aire de la distribution prédite avec le grain.

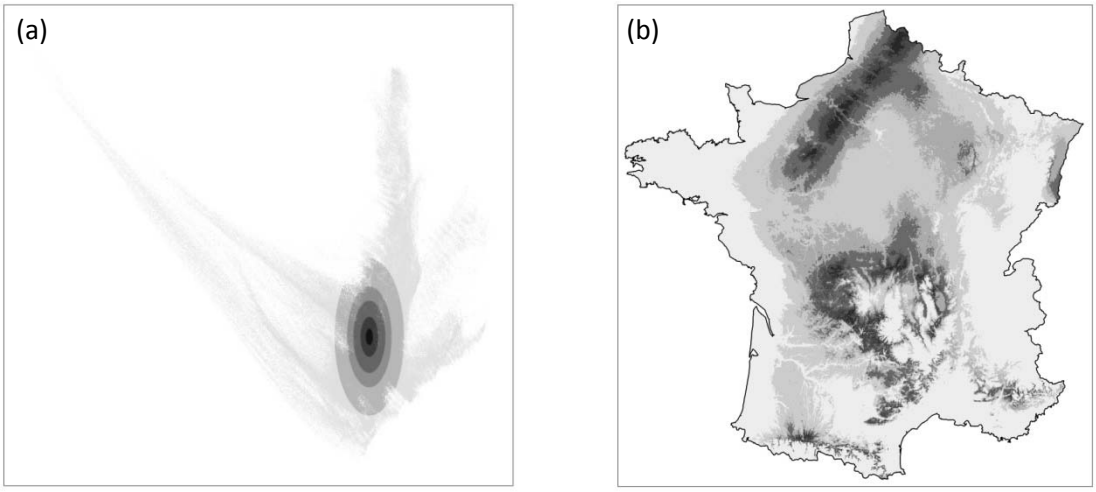


Figure 11 : Les niches environnementales dans le plan défini par les deux variables environnementales synthétiques (a) et les distributions sur le territoire français (b) des espèces virtuelles utilisées dans les articles M1 et M3. Chaque niche représentée par un niveau de gris contient les niches plus petites représentées par des niveaux de gris plus foncés. Les cinq espèces correspondent aux prévalences 1%, 5%, 15%, 30% et 60%.

Le but de notre première étude a été d'expliquer cet accroissement, ce qui a permis de montrer que les résultats obtenus par Guisan *et al.* (2007) d'une part, et Seo *et al.* (2009) et Hu & Jiang (2010) d'autre part n'étaient en fait pas contradictoires.

Pour simplifier l'étude et mettre en évidence l'origine méthodologique du phénomène nous avons décidé de travailler sur des espèces virtuelles construites à partir de données climatologiques réelles sur le territoire français. Nous avons sélectionné 8 variables

climatiques, puis réalisé une ACP pour garder les deux premiers axes comme variables climatiques synthétiques. Les niches de nos espèces virtuelles ont été définies comme des disques de l'espace de ces deux variables. La distribution d'une espèce est l'ensemble des carrés (pixels) pour lesquels le couple des deux variables synthétiques est inclus dans la niche (Figure 11). Le grain de travail le plus fin est celui des données climatologiques Worldclim (Hijmans et al. 2005) utilisées, soit des pixels d'environ 1 km x 1 km. Nous avons utilisé une approche de modèle d'ensemble et nous avons comparé les projections obtenues au grain initial et à six autres grains, le passage d'un grain au suivant se faisant par agrégation de 4 pixels adjacents.

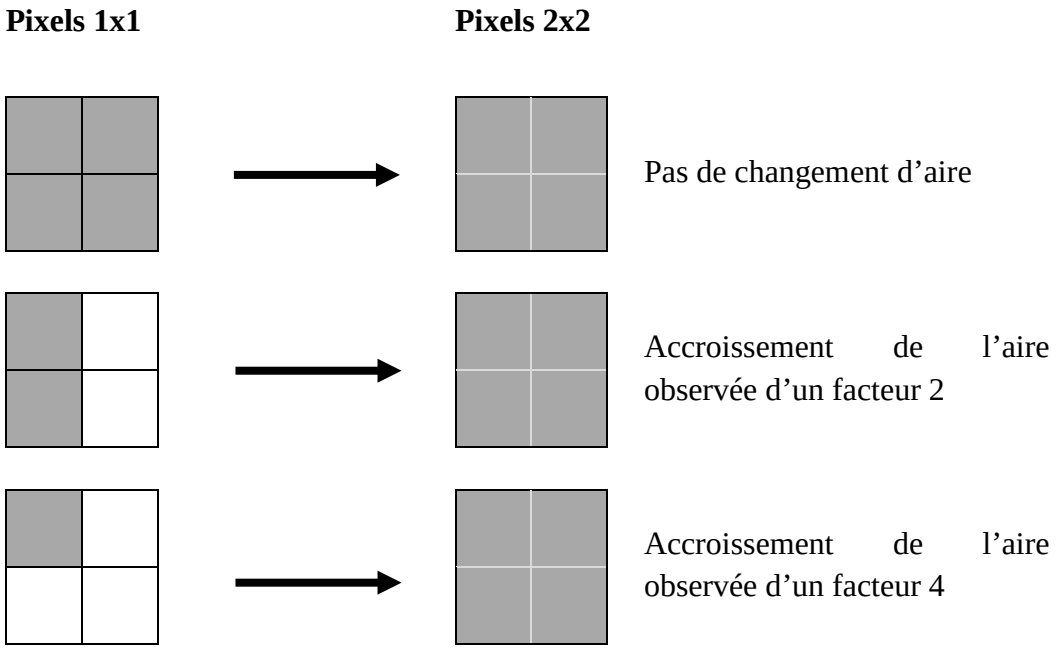


Figure 12 : Accroissement « géométrique » de l'aire observée par changement de grain. La zone grisée correspond aux pixels où l'espèce est présente, alors que l'espèce est absente dans les pixels non grisés.

Nous avons montré que l'accroissement de l'aire prédite par les modèles de niches était essentiellement dû à l'accroissement géométrique de l'aire observée (Figures 12, 13). Cet effet géométrique doit être pris en compte lors de l'évaluation de la surface des niches, en

particulier dans le cas d'espèces à niche très fragmentée, pour lesquelles il sera majoré. Ces résultats sont donc importants dans le cadre de plans de protection, les espèces en danger ayant souvent une distribution fragmentée liée à la perte de leur habitat. L'accroissement de l'aire prédite n'étant pas dû à la modélisation, il ne traduit pas une perte de qualité des modèles. D'ailleurs, les différents indices utilisés pour mesurer la qualité des projections n'ont mis en évidence une baisse significative de la qualité que lors de l'utilisation de grains très larges, qui conduit à une perte importante d'information environnementale.

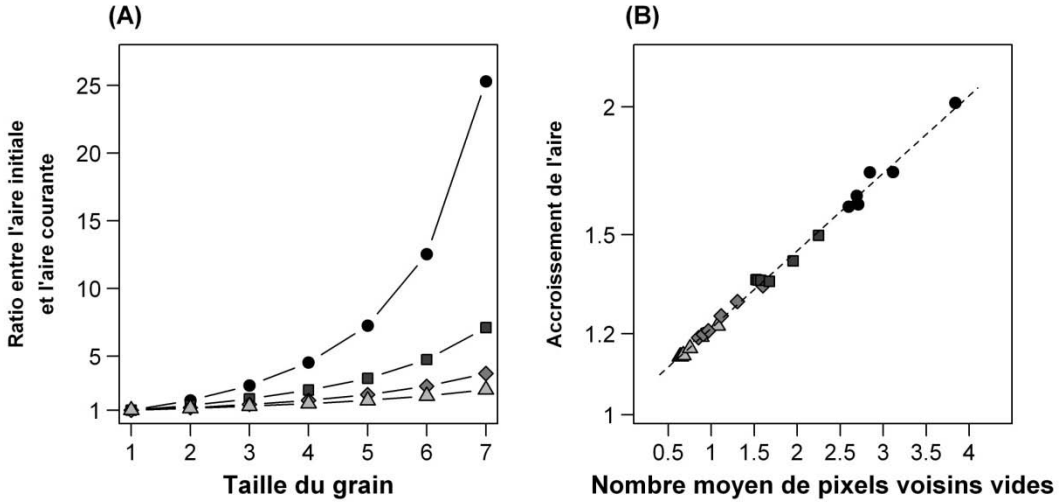


Figure 13 : Effet de la taille du grain sur l'aire de la distribution observée. (A) Ratio entre l'aire de la distribution observée au grain le plus fin **(aire initiale)** et l'aire mesurée au grain considéré **(aire courante)**. Les aires sont mesurées comme le nombre de pixels au grain le plus fin contenus dans la distribution. (B) Lien entre le nombre de pixels vides autour de la distribution et l'accroissement de l'aire observé en passant d'un grain au suivant. Les symboles représentent la prévalece de l'espèce. Cercles noirs 1% ; carrés gris foncés : 5% ; losanges gris : 15% ; triangles gris pâle : 30%.

Cependant, le changement de grain entraine des modifications dans la répartition géographique des erreurs, les modèles construits au grain le plus fin omettant parfois de larges zones géographiques alors que pour des grains un peu plus gros, les fausses absences sont réparties beaucoup plus uniformément en bord de distribution (Figure 14). Le compromis nécessaire entre la précision (grain) et l'étendue (amplitude des conditions environnementales échantillonnées) des données (Braunisch and Suchant 2010), l'augmentation de l'incertitude sur les données climatiques liées à la topographie et à la couverture végétale (Wiens and

Bachelet 2009; Dobrowski 2011) et la différence de répartition géographique des erreurs peuvent conduire à préférer un grain qui ne soit pas minimal. Le choix de l'échelle et des variables utilisées doit aussi prendre en compte les processus écologiques impliqués (Austin and Van Niel 2011) et l'objectif de l'étude (Lomba et al. 2010). Dans certains cas, on peut également envisager l'utilisation simultanée de différentes échelles (Mateo-Tomas and Olea 2009; Huber et al. 2010).

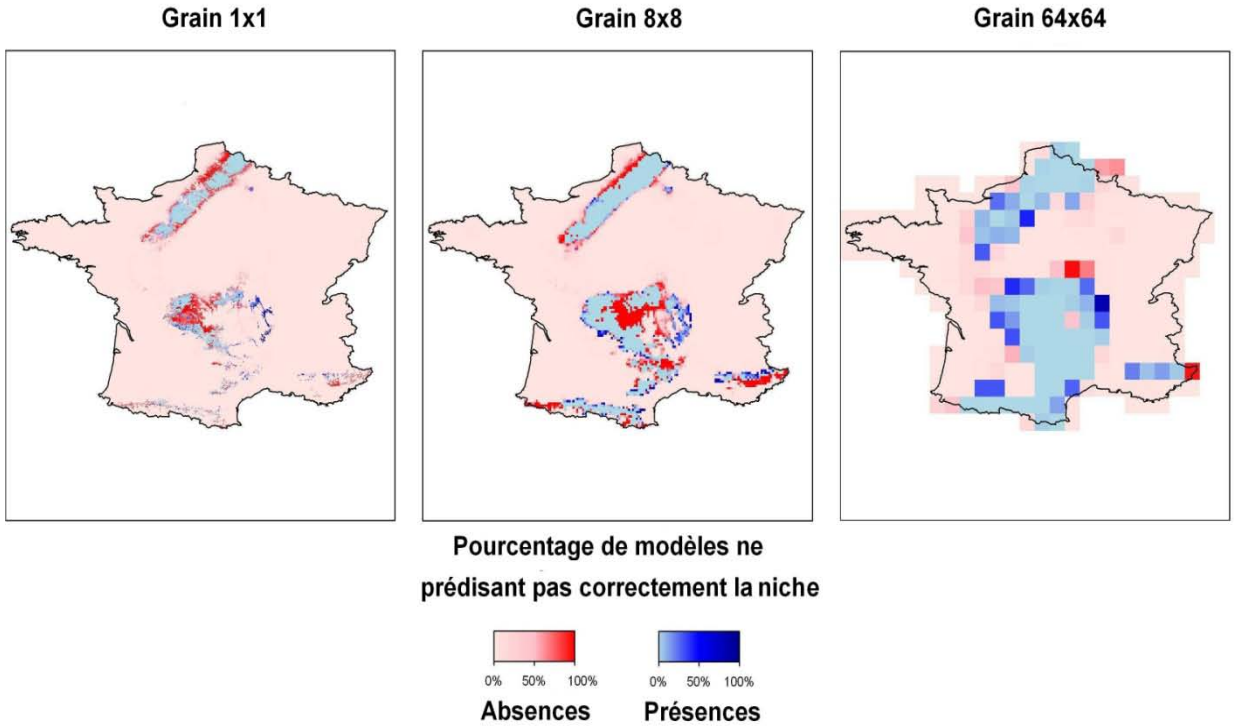


Figure 14 : Effet du grain sur la distribution prédite pour l'espèce de prévalence 5%. On remarque que toute la bordure est du Massif central est omise par les modèles au grain le plus fin (plus un pixel est sombre, plus le nombre de modèles le prédisant incorrectement est élevé).

Cette étude a mis en évidence que :

- **le choix du grain est capital dans l'évaluation de la surface de la distribution (observée ou prédite) en particulier dans le cas de distributions fragmentées ;**
- **le grain, tant qu'il reste dans une gamme raisonnable, a peu d'influence sur la qualité des modèles ;**

- **il n'existe probablement pas de choix idéal de taille de grain pour la construction des modèles. Le grain doit être sélectionné en prenant en compte les objectifs de l'étude, les caractéristiques de l'espèce et la qualité des données.**

B. L'utilisation des données à large échelle pour identifier les changements de niche (M2)

Lorsque les données sont mesurées à un grain très grossier, le bassin versant ou le pays par exemple, il devient impossible de les utiliser pour construire des modèles de distribution corrélatifs. Le but de notre deuxième étude a été de tester si ces données à un grain très large restaient cependant utilisables pour mettre en évidence des changements de niche dans le cadre de l'étude des espèces invasives. Les nombreuses études portant sur le changement de niche entre la zone native et la zone exotique ayant abouti à des résultats parfois contradictoires, il est important de pouvoir tester cette hypothèse sur le plus grand nombre d'espèces possibles. La possibilité d'utiliser des données d'atlas ou d'inventaires régionaux, disponibles pour de très nombreuses espèces, répond tout à fait à ce besoin.

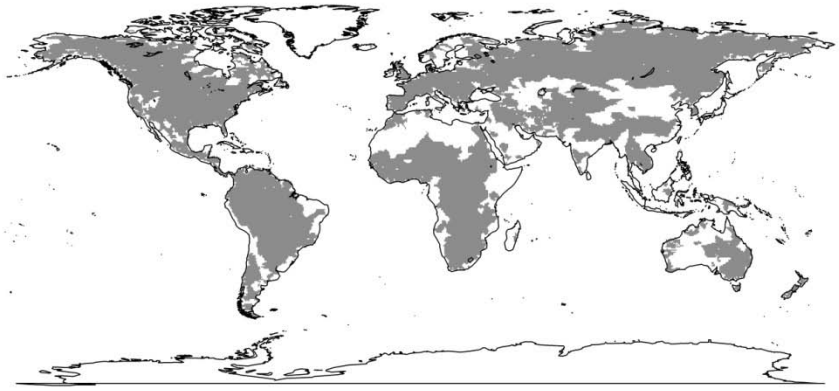


Figure 15 : Les 1055 bassins (Brosse *et al*, Annexe 1) utilisés dans l'étude sur l'utilisation des données à grain large pour l'identification des changements de niche (M2).

Pour cette étude, nous avons utilisé une base de données de présence/absence de poissons à l'échelle mondiale pour 1055 bassins couvrant plus de 80% de la surface continentale (Figure 15).

Espèces	Occurences natives	Occurences exotiques	% de bassins de l'aire exotique avec un shift
Ameiurus melas	18	49	14.3
Carassius auratus	49	164	12.8
Carassius carassius	30	54	55.6
Cyprinus carpio	34	245	20.4
Gambusia sp.	52	206	20.4
Ictalurus punctatus	33	44	25.0
Lepomis cyanellus	20	46	23.9
Lepomis gibbosus	28	52	23.1
Lepomis macrochirus	44	63	19.0
Micropterus salmoides	48	100	24.0
Oncorhynchus mykiss	56	189	27.5
Perca fluviatilis	103	64	42.2
Pseudorasbora parva	29	45	15.6
Salmo trutta	141	131	21.4
Salvelinus fontinalis	66	71	40.8
Sander lucioperca	32	48	43.8
Thymallus thymallus	35	33	12.1
Tinca tinca	79	46	10.9

Tableau 1 : Nombre d'occurrences natives et non natives des 18 espèces étudiées et pourcentage des bassins pour lesquels au moins un changement de niche a été observé pour l'une des variables considérées.

A l'intérieur de chaque bassin, les zones de présence de l'espèce n'étant pas connues, il n'était pas possible de déterminer pour chaque espèce les conditions environnementales précises de sa niche. La surface des bassins présentait de grandes disparités, puisque la base contient aussi bien des données sur le bassin de l'Amazone et du Mississipi que celles de petits fleuves côtiers. La majorité des espèces n'habitant qu'une partie des bassins où elles sont présentes, la valeur moyenne d'une variable climatique, calculée sur l'ensemble du bassin, peut être très éloignée des conditions réellement adaptées à l'espèce. Nous avons donc fait le choix de supposer que toutes les conditions environnementales rencontrées dans le bassin étaient potentiellement adaptées à l'espèce ciblée. Si cette hypothèse est clairement irréaliste dans le

cas des très grands bassins et peut conduire à une non-identification de changement de niche, elle présente l'avantage de garantir que les changements de niche identifiés à l'échelle du bassin sont réels, un changement de niche étant identifié lorsque l'intersection entre les conditions environnementales de l'aire native et de l'aire exotique est vide.

Nous avons sélectionné 18 espèces de poissons largement introduites dans le monde (Tableau 1) et pour lesquelles nous disposions d'un nombre suffisant d'occurrences natives afin de ne pas identifier de changement de niche à cause d'un échantillonnage insuffisant des conditions de l'aire native. Les variables environnementales ont été choisies parmi les variables pertinentes dans la description des niches environnementales de poissons et utilisées dans de nombreuses études (Minns and Moore 1995; Chu et al. 2005; Leprieur et al. 2009b) : précipitations du mois le plus sec, précipitations du mois le plus humide, coefficient de variation des précipitations mensuelles, nombre de jours de pluie, température moyenne du mois le plus froid, température moyenne du mois le plus chaud, coefficient de variation de la température mensuelle moyenne, amplitude moyenne des températures.

Nous avons montré que toutes les espèces présentaient un changement de niche pour au moins une variable. Ce résultat pourrait s'expliquer par le choix de variables non pertinentes pour la distribution des poissons. Cela pourrait éventuellement être le cas pour le coefficient de variation de la température mensuelle moyenne, qui montre un changement pour plus de trois quart des espèces et un quart des bassins. Mais cela ne peut en aucun cas expliquer l'ensemble de nos résultats, la truite arc en ciel (*Oncorhynchus mykiss*), par exemple, montrant un changement de niche pour toutes les variables (Figure 16) et pour un quart de ses bassins. Une telle tendance au changement de niche chez les poissons d'eau douce, contrairement à ce qu'on observe dans d'autres taxons (Petitpierre et al. 2012), est probablement liée à la nature fragmentée des milieux d'eau douce qui limite très fortement la dispersion des espèces. Les conditions environnementales rencontrées dans l'aire native traduisent plus des contraintes

historiques et géographiques que des limitations physiologiques. La truite arc en ciel, originaire d'une petite partie de l'Amérique du Nord, a été introduite dans très nombreuses régions, et souvent avec une forte pression de propagule. Elle a donc eu la possibilité de s'établir dans de nombreux lieux adaptés à sa physiologie que les contraintes de dispersion lui avaient jusque là interdit. Cela explique le nombre très important de changements de niches observés. Le phénomène peut avoir été amplifié par des adaptations : plasticité, mutation (e.g., Meffe 1991; Meffe et al. 1995).

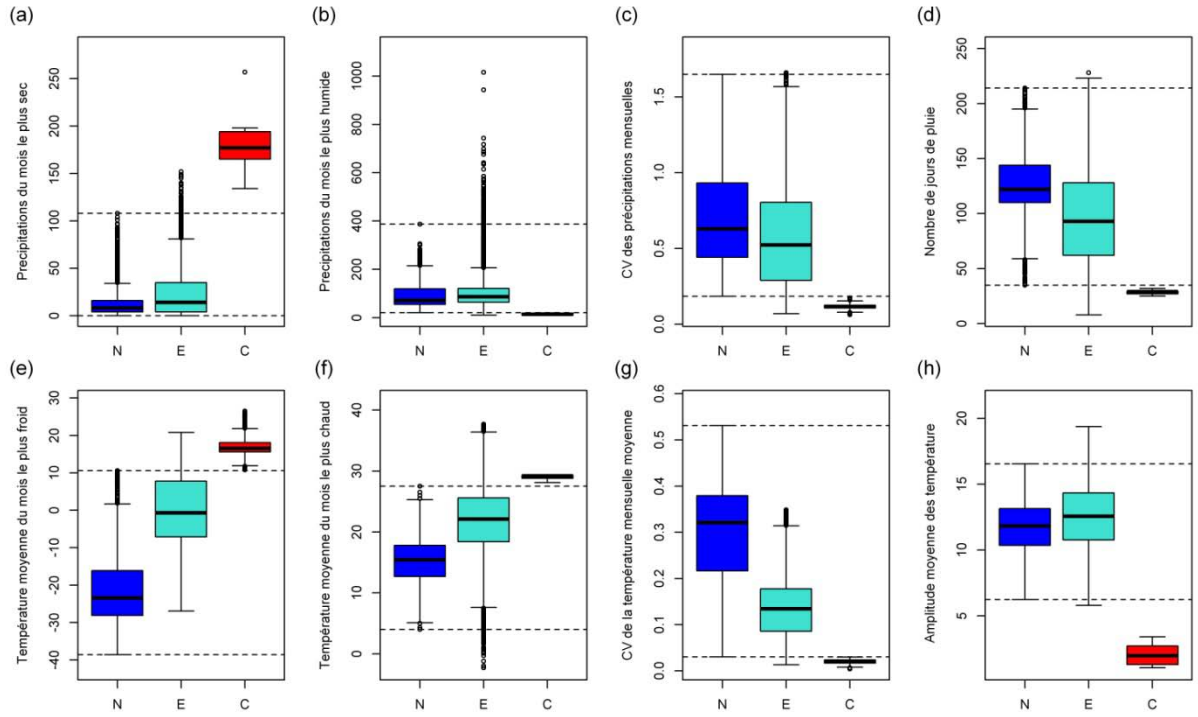


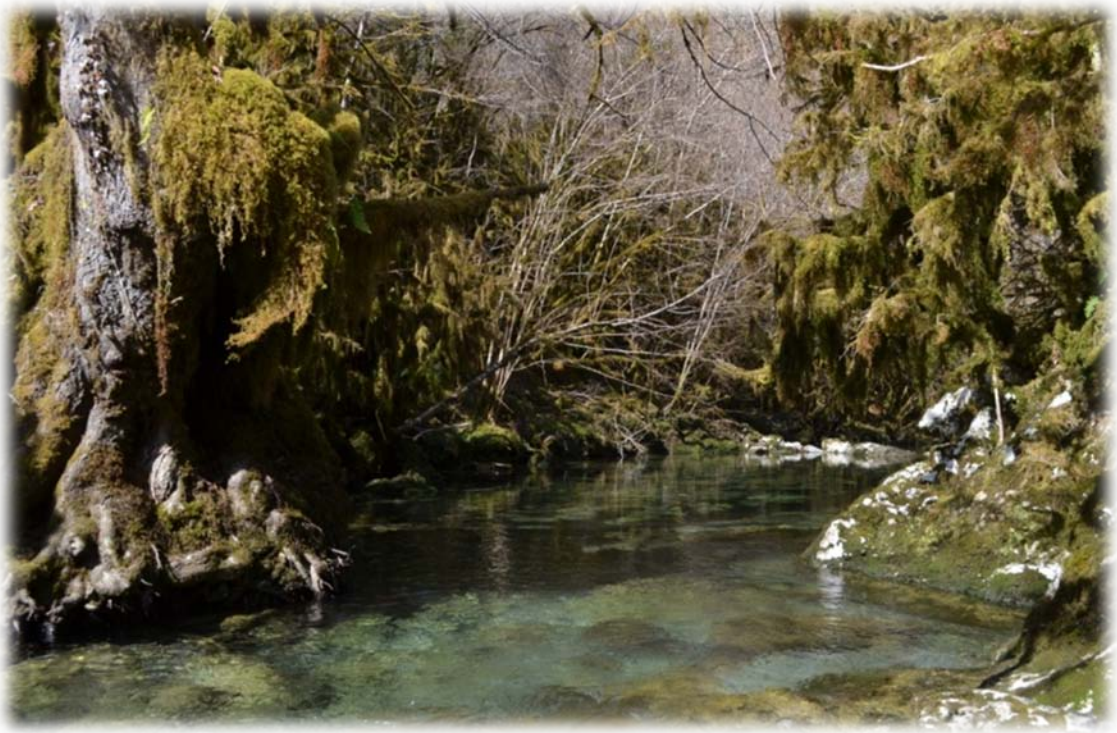
Figure 16 : Changement de niche observé pour la truite arc en ciel pour les 8 variables climatiques considérées : (a) précipitations du mois le plus sec, (b) précipitations du mois le plus humide, (c) coefficient de variation des précipitations mensuelles, (d) nombre de jours de pluie, (e) température moyenne du mois le plus froid, (f) température moyenne du mois le plus chaud, (g) coefficient de variation de la température mensuelle moyenne, (h) amplitude moyenne des températures. Les trois boites à moustaches représentent l'ensemble des conditions climatiques rencontrées dans les bassins : de l'aire native (bleu) ; de l'aire exotique et ne présentant pas de changement (turquoise) ; de l'aire exotique et présentant un changement (rouge).

Les changements de niche relativement rares d'autres espèces peuvent eux aussi avoir différentes explications : introduction essentiellement dans des milieux où l'environnement climatique est similaire à celui de l'aire native, espèce dont la distribution native couvre la quasi-totalité de la gamme des conditions environnementales adaptées à sa physiologie, grain

très grossier de cette étude (en particulier pour les espèces présentes surtout dans de grands bassins)… Le grain utilisé ici ne permet pas de répondre à ces questions mais suffit pour identifier certaines espèces susceptibles de faire l'objet d'une étude approfondie pour étudier les causes des changements de niche et leur amplitude réelle.

Cette étude a mis en évidence que :

- **le changement de niche semble être un phénomène répandu chez les poissons d'eau douce, contrairement à ce qui a été observé dans d'autres taxons ;**
- **des données à très large échelle peuvent être utilisées pour identifier des changements de niche et cibler les espèces à privilégier pour une étude plus détaillée.**



Partie II :
Les risques d'établissement d'espèces destinées à l'aquaculture sur le territoire français

A. Introduction

Les poissons sont une source importante de protéines dans le monde, mais les stocks disponibles sont en déclin suite à la destruction des habitats, à l'introduction d'espèces non-natives, à la pollution et à la surexploitation (Millennium Ecosystem Assessment 2005). Si les premières réglementations contraignantes visant à limiter la surexploitation sont apparues dès le début des années 70, avec entre autre les quotas de pêche à l'anchois, elles ont porté quasi-exclusivement sur les espèces marines. Ce n'est que depuis le début des années 2000 que la situation des eaux douces fait l'objet d'études à large échelle. Ces études ont mis en évidence que les milieux d'eau douce devaient eux aussi faire face à une pression de pêche croissante dans les pays en développement, avec une augmentation depuis 50 ans d'environ 3% par an des quantités prélevées (Allan et al. 2005). Dans les pays développés, la tendance est inverse, les pêches commerciales étant progressivement remplacées par des pêches de loisir. La différence entre pays développés et pays en développement est également notable pour l'aquaculture. Si la production mondiale annuelle a plus que triplé ces vingt dernières années (source : FAO 2010), la production européenne est à la baisse depuis 10 ans (source : Eurostat). La demande se fait donc de plus en plus pressante de la part des aquaculteurs pour obtenir l'autorisation d'introduire sur le territoire européen de nouvelles espèces, choisies en particulier pour leur croissance rapide. Mais l'aquaculture est une source majeure d'espèces invasives (Casal 2006), avec généralement une pression de propagule élevée, ce qui favorise le processus d'établissement (Ruesink 2005). Il est donc crucial d'évaluer le risque d'établissement potentiel des espèces faisant l'objet d'une demande d'introduction pour l'aquaculture, non seulement dans les conditions actuelles mais aussi en tenant compte des conditions climatiques futures prédites par les différents modèles climatiques pour les différents scénarios d'émission de gaz à effet de serre.

Le règlement européen (nº 708/2007) relatif à l'utilisation en aquaculture des espèces non natives liste un certain nombre d'espèces pouvant être introduites sans étude d'impact préalable, sauf si certains états membres souhaitent restreindre l'usage de ces espèces sur leur territoire. Il concerne en particulier les carpes asiatiques (*Ctenopharyntgodon idella, Hypophthalmichthys molitrix, Hypophthalmichthys nobilis),* autorisées à l'introduction en eau close en France jusqu'en 2006, en particulier pour le contrôle de la prolifération des végétaux dans les bassins d'aquaculture. Leur reproduction en milieu naturel, qui nécessite des conditions hydrologiques particulières et des températures élevées (Shireman and Smith 1983), n'a pas été observée sur le territoire français jusqu'à ce jour mais le réchauffement climatique pourrait favoriser l'établissement de ces espèces. Le règlement concerne aussi l'achigan à grande bouche ou black-bass (*Micropterus salmoides*). Cette dernière espèce, originaire du continent nord américain, est un prédateur vorace qui impacte fortement les écosystèmes receveurs (Shelton et al. 2008; Weyl et al. 2010). Elle a été très largement introduite pour la pêche sportive et est maintenant établie en de nombreux points du territoire français.

Les demandes d'introduction pour l'aquaculture portent également sur deux espèces de Tilapia (originaires d'Afrique) : *Oreochromis mossambicus* et *Oreochromis niloticus*. Leur élevage se ferait surtout, au moins dans un premier temps, dans les DOM, dont les conditions climatiques sont proches de celles de leur aire native. L'un d'eux, *Oreochromis mossambicus,* est d'ailleurs déjà présent dans certains des départements d'Outre Mer (source : ISSG (IUCN)). Deux espèces de Siluridés, l'une américaine (*Ictalurus punctatus*), l'autre africaine (*Clarias gariepinus*), sont également concernées.

L'Office National de l'Eau et des Milieux Aquatiques (ONEMA) est un organisme public chargé de l'étude et de la surveillance de l'état des eaux et du fonctionnement des écosystèmes dulçaquicoles dans le but de favoriser une gestion globale et durable de la

ressource en eau et des écosystèmes aquatiques. Dans le cadre de son rôle de gestion durable des milieux aquatiques continentaux, l'ONEMA a souhaité connaître les risques d'établissement de cinq espèces (Tableau 2) faisant l'objet de demandes d'introductions pour l'aquaculture ainsi que de *Micropterus salmoides*, déjà présent en France et concerné par la circulaire européenne n° 708/2007. L'ONEMA effectuant un suivi régulier des populations piscicoles du réseau français en organisant des campagnes de pêche électrique, on dispose de plus de 150 occurrences de *Micropterus salmoides* en France métropolitaine. L'étude de cette espèce présente donc un double intérêt :

- prédiction des zones potentielles d'établissement en dehors de l'aire déjà colonisée et sous l'effet du changement climatique ;
- comparaison de la distribution prédite par les modèles construits sur l'aire native avec la distribution observée pour évaluer la qualité des modèles et mettre en évidence un éventuel changement de niche.

Espèce	Ordre	Nom commun	Continent	Intérêt	Occurrences
Micropterus salmoides	Perciforme	Achigan à grande bouche	Amérique du nord	Pêche sportive, aquaculture	386 (USA) 152 (France)
Ictalurus punctatus	Siluriforme	Barbue d'Amérique	Amérique du nord	Aquaculture	200
Clarias gariepinus	Siluriforme	Poisson chat africain	Afrique	Aquaculture	129
Oreochromis mossambicus	Cichlidé	Tilapia du Mozambique	Afrique	Aquaculture	143
Oreochromis niloticus	Cichlidé	Tilapia du Nil	Afrique	Aquaculture	75
Ctenopharyngodon idella	Cyprinidés	Amour blanc	Asie	Aquaculture, lutte biologique	82

Tableau 2 : Espèces étudiées dans le cadre du projet INVAQUA, raisons de leur introduction sur le territoire français et nombre d'occurrences utilisées dans la construction des modèles.

Dans le but d'évaluer les risques d'établissement de ces six espèces, l'ONEMA a mis en place le programme INVAQUA (INVAsion des milieux AQUAtiques) en collaboration avec le Muséum d'Histoire Naturelle de Paris et le laboratoire Evolution et Diversité Biologique de Toulouse. Dans ce cadre, nous avons été chargés d'établir les cartes de distribution potentielles des espèces d'intérêt pour la période actuelle et sous l'effet du changement climatique en France métropolitaine et dans les DOM.

B. Matériel et méthode

1. Les données environnementales

Deux types de variables environnementales ont été utilisés pour modéliser la distribution des espèces étudiées. Nous avons tout d'abord sélectionné deux variables topographiques disponibles à large échelle : l'accumulation de flux et la pente. Ces deux variables sont issues du modèle numérique de terrain SRTM30 plus et permettent d'évaluer la position du site dans le gradient amont-aval ainsi que sa turbulence. Nous avons également utilisé 4 variables climatiques, choisies pour leur pertinence dans la description de la niche des poissons et sélectionnées pour limiter les corrélations entre variables. Nous avons retenu les températures moyennes de l'air du trimestre le plus chaud et du trimestre le plus froid ainsi que les précipitations du trimestre le plus humide et du trimestre le plus sec. Les données climatiques actuelles sont les données Worlclim (Hijmans et al. 2005). Les données futures utilisées sont dérivées de deux modèles de circulation CGCM (Canadian Centre for Climate Modelling and Analysis) et HadCM (Hadley Centre for Climate Prediction and Research's General Circulation Model) pour le scénario A1B et un modèle (HadCM) pour le scénario B2A. Toutes les variables utilisées sont à l'échelle 30"x30".

2. Les données d'occurrence

Les données d'occurrence des espèces américaines proviennent de campagnes d'échantillonnage sur l'ensemble du territoire des Etats-Unis (Mitchell and Knouft 2009) dans le cadre du programme National Water Quality Assessment (NAWQA). Ce programme a fait l'objet à l'échelle nationale de processus standardisés de collectes de données. Il a pour but de suivre l'évolution de la qualité de l'eau au cours du temps et l'impact de l'environnement et des activités humaines sur cette qualité. Les communautés de poissons ont été échantillonnées dans près de 1000 sites, mais les données d'absence des espèces sont peu fiables en termes de niche climatique et environnementale. En effet, le programme NAWQA ayant pour but d'évaluer la qualité des eaux, plusieurs des sites suivis sont très fortement perturbés. L'absence des espèces n'est alors pas liée aux conditions climatiques et topographiques mais due aux perturbations.

Figure 17 : Occurrences des deux espèces américaines.

Les données d'occurrence de *Ctenopharyngodon idella* proviennent de deux sources : celles de la zone native sont issues de DeVaney et al. (2009) ; celles de la zone exotique (USA) de cette même publication ainsi que de la base américaine. Nous ne disposions d'aucune donnée d'absence sur la zone native.

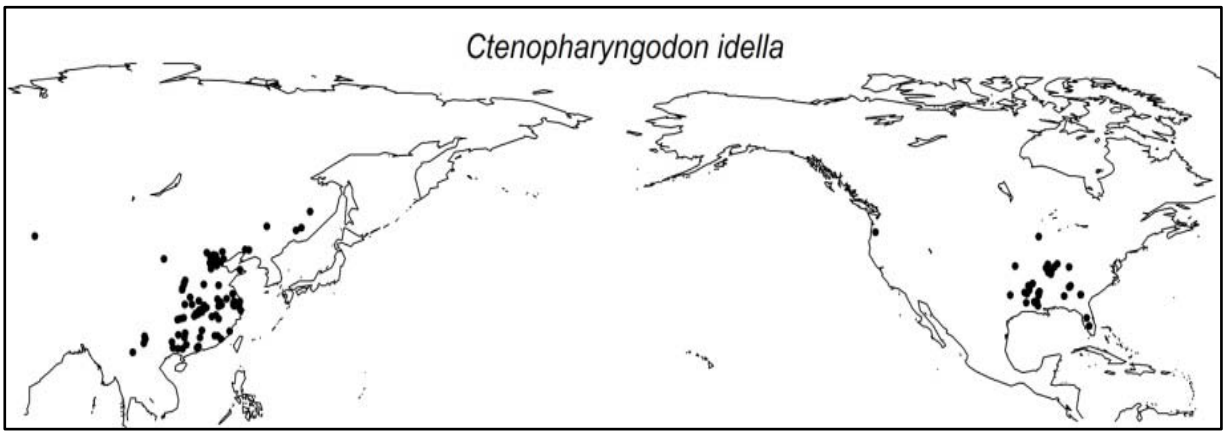


Figure 18 : Occurrences de l'espèce asiatique.

Les données d'occurrence des trois espèces africaines proviennent de la base de données Faunafri développée par l'IRD (Institut de Recherche pour le Développement). Celle-ci regroupe les occurrences observées lors d'inventaires destinés principalement à des collections de Muséum National d'Histoire Naturelle de 1850 à nos jours. Les données bioclimatiques concernant la période actuelle utilisées dans cette étude sont une synthèse sur la période 1960-1990. Nous avons donc sélectionné les seules données d'occurrence collectées pendant cette période. Les échantillonnages ayant servi à alimenter la base Faunafri étant souvent ciblés sur une ou quelques espèces d'intérêt, les données d'absence sont peu représentatives d'une absence réelle de l'espèce.

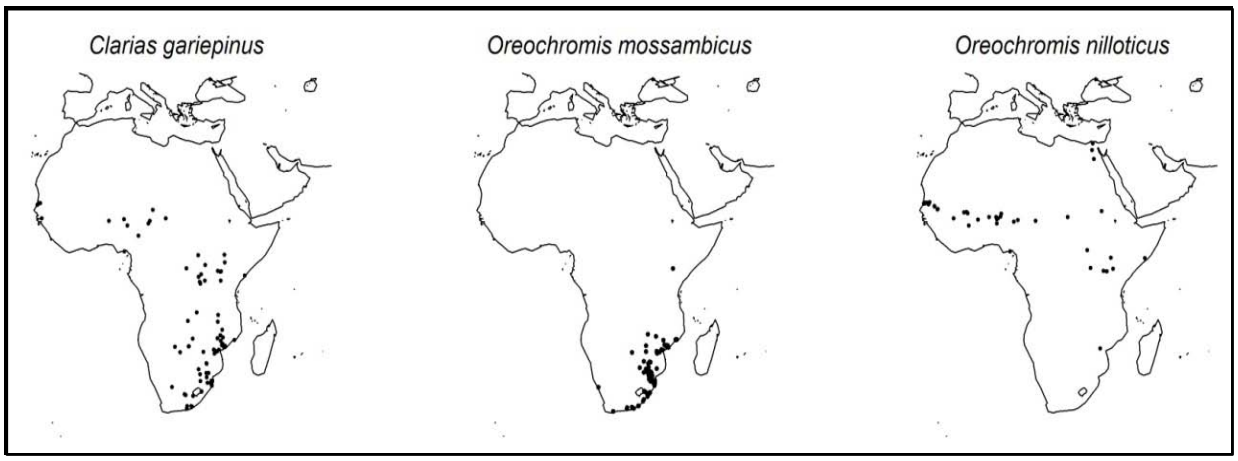


Figure 19 : Occurrences des trois espèces africaines.

Pour les six espèces, les données d'absence étaient soit indisponibles soit peu fiables. Nous avons donc fait le choix de sélectionner aléatoirement des pseudo-absences parmi l'ensemble des points du réseau hydrographique de la région concernée et situés à au moins 3° d'un site d'observation de l'espèce.

3. Modélisation

La modélisation de niche s'est faite en utilisant une approche d'ensemble classique utilisant six méthodes statistiques: les modèles linéaires généralisés (GLM) ; les modèles additifs généralisés (GAM); les boosted trees (BT); les arbres de régression et de classification (CART); les generalized boosted regression models (GBM) et l'analyse linéaire discriminante (LDA). Pour les modèles linéaires généralisés et l'analyse linéaire discriminante, les variables au carré ont été intégrées dans le modèle pour tenir compte de la non linéarité des réponses biologiques.

Les modèles générés ont permis d'obtenir 6 probabilités de présence qui ont été moyennées (Marmion et al. 2009) pour obtenir une probabilité de présence unique convertie en données de présence/absence par une valeur seuil déterminée en maximisant la TSS. Les 6 méthodes ont été calibrées sur 70% des sites (présences et pseudo-absences) disponibles, leur qualité évaluée sur les 30% restants. La qualité des modèles a été évaluée en utilisant une mesure indépendante de la valeur seuil : l'AUC et deux mesures dépendantes de la valeur seuil : le Kappa et la TSS.

Les modèles générés ont ensuite été utilisés pour prédire les distributions potentielles actuelles et futures sur l'ensemble du réseau hydrographique français. Le processus a été répété 100 fois i.e. sur 100 choix aléatoires de pseudo-absences (en nombre égal aux présences) et divisions aléatoires des données en base d'apprentissage et base de test.

Afin de mieux évaluer la fiabilité de nos modèles et d'estimer si *Micropterus salmoides* avait changé de niche, nous avons également utilisé les occurrences françaises de deux façons différentes. Nous avons tout d'abord regardé le niveau de risque prédit pour les points de présence sur le territoire français. Nous avons ensuite utilisé les données d'occurrence en France pour prédire la distribution de l'espèce à la fois en France et aux USA et nous avons comparé ces prédictions avec celles obtenues à partir des données d'occurrence américaine (Medley 2010; Hill et al. 2012). Le processus de modélisation a été le même que pour les autres cas. La seule différence réside dans le choix des pseudo-absences qui ont été prises à plus d'un demi degré des points de présence (au lieu de 3°) à cause de la faible étendue de la zone géographique considérée.

C. Résultats

1. Résultats généraux

	Micropterus salmoides (USA)	***Micropterus salmoides (France)***	***Ictalurus punctatus***	***Clarias gariepinus***	***Oreochromis mossambicus***	***Oreochromis nilloticus***	***Ctenopharyngodon idella***
AUC	0.983±0.005 (min 0.969)	0.921±0.027 (min 0.811)	0.984±0.007 (min 0.968)	0.881±0.04 (min 0.748)	0.98±0.013 (min 0.944)	0.841±0.058 (min 0.724)	0.966±0.027 (min 0.865)
TSS	0.933±0.014 (min 0.9)	0.847±0.077 (min 0.707)	0.934±0.021 (min 0.88)	0.793±0.045 (min 0.636)	0.926±0.035 (min 0.823)	0.768±0.073 (min 0.592)	0.915±0.041 (min 0.82)
Kappa	0.866±0.029 (min 0.799)	0.692±0.038 (min 0.410)	0.869±0.043 (min 0.748)	0.584±0.09 (min 0.27)	0.851±0.071 (min 0.635)	0.533±0.148 (min 0.129)	0.828±0.083 (min 0.602)

Tableau 3 : Qualité des modèles obtenus pour les 6 espèces considérées.

Les modèles obtenus sont de bonne, voire très bonne qualité (Tableau 3). Seules deux espèces (*C. gariepinus et O. niloticus*) ont des modèles de performance parfois réduite avec des valeurs d'AUC inférieures à 0.8 et de Kappa inférieures à 0.4. On note également une moindre qualité des modèles de *M. salmoides* construits à partir de la base française. Les patrons des risques d'établissement diffèrent suivant les espèces et les régions (métropole ou DOM) considérées. Mais les prévisions semblent relativement peu affectées par les scénarios

d'émission de gaz à effet de serre considérés. Ces derniers interviennent surtout dans l'amplitude des modifications de distribution observées.

2. Risques d'établissement en métropole

a) Micropterus salmoides

Cette espèce est déjà présente en de nombreux points du territoire puisque elle a été observée dans 152 sites suivis par l'Onema. Dans les conditions actuelles, elle serait capable de coloniser l'ensemble du bassin Adour-Garonne et les bassins du Rhône et de la Loire mais aussi les cours d'eau principaux du bassin de la Seine et le pourtour méditerranéen, soit près de 50% du réseau hydrographique métropolitain. Ces prédictions correspondent au patron de distribution des présences observées (nombreuses observations en bassin Adour-Garonne qui est la zone connaissant les risques maximaux, autres observations dans les autres bassins à risque). Seuls 3 sites où l'espèce a été observée sont prédits comme à risque par moins de 25% des modèles (l'espèce n'ayant été observée qu'une seule fois sur ces sites), un tiers des sites de présence sont prédits par tous les modèles et 88% des sites de présence sont prédits par plus de 50% des modèles. Cela semble indiquer une bonne concordance entre niche native et niche exotique ainsi qu'une bonne qualité des modèles.

Ce résultat est confirmé par la comparaison des prédictions de distribution en France métropolitaine obtenues avec les deux bases. On obtient deux distributions géographiquement proches (Figure 21 a-b) et la différence entre les deux niveaux de risque reste généralement modérée (inférieure à 30% en valeur absolue dans plus de 77% des cas). Les risques prédits en utilisant les occurrences françaises sont généralement inférieurs avec 25 % des sites pour lesquels le niveau de risque est plus élevé en utilisant les occurrences françaises et seulement 7% pour lesquels la différence est supérieure à 10% (Figure 21 c). Ces sites, situés majoritairement dans le bassin de la Loire et le cours moyen de la Seine, sont généralement

déjà prédits comme à risque par les modèles utilisant les occurrences de l'aire native (plus de 90% de ces sites pour lesquels plus de 30% des modèles basés sur les données américaines prédisent l'établissement).

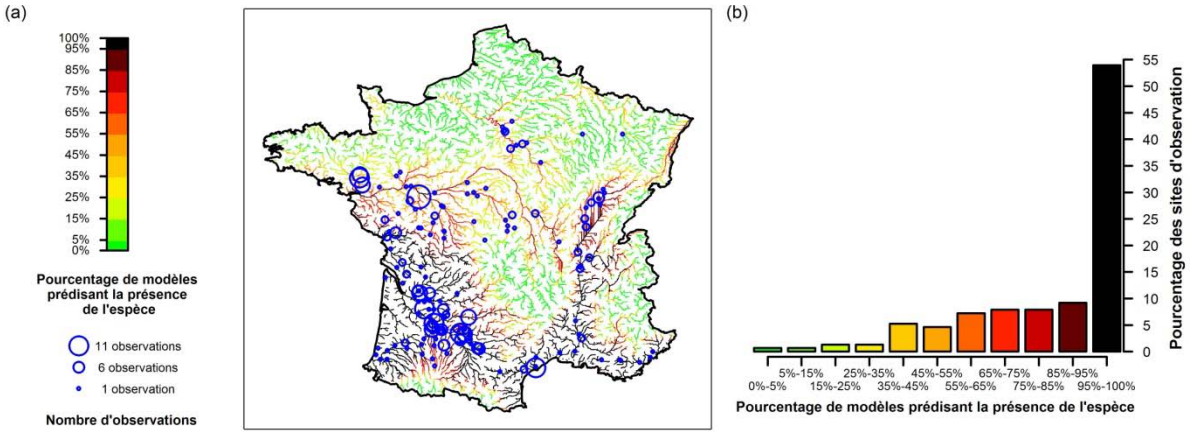


Figure 20 : Risque d'établissement de *Micropterus salmoides* en métropole sous les conditions environnementales actuelles. (a) Niveau de risque sur l'ensemble du réseau hydrographique métropolitain et points d'observation de l'espèce. Le risque correspond au pourcentage de modèles prédisant l'espèce comme potentiellement présente à un point donné. Les cercles bleus correspondent aux observations de l'espèce. La taille des cercles traduit le nombre d'observations de l'espèce sur ce site. (b) Niveau de risque des 152 sites d'observation de l'espèce.

Par contre, on note une très grande différence de prédiction de distribution aux USA entre les deux bases. Avec l'utilisation de la base américaine, toute la moitié est des Etats-Unis ainsi que la bordure ouest des Rocheuses sont prédites comme zone de distribution potentielle de l'espèce. On notera que la quinzaine d'occurrences du centre du pays se trouve dans des zones non prédites par les modèles. L'utilisation de la base française, quant à elle, ne fait apparaître comme zone d'habitat potentiel que quelques régions de l'ouest des Rocheuses ainsi que, dans une moindre mesure, le pourtour du golfe du Mexique, la grande majorité des occurrences de l'espèce se trouvant en dehors des zones prédites.

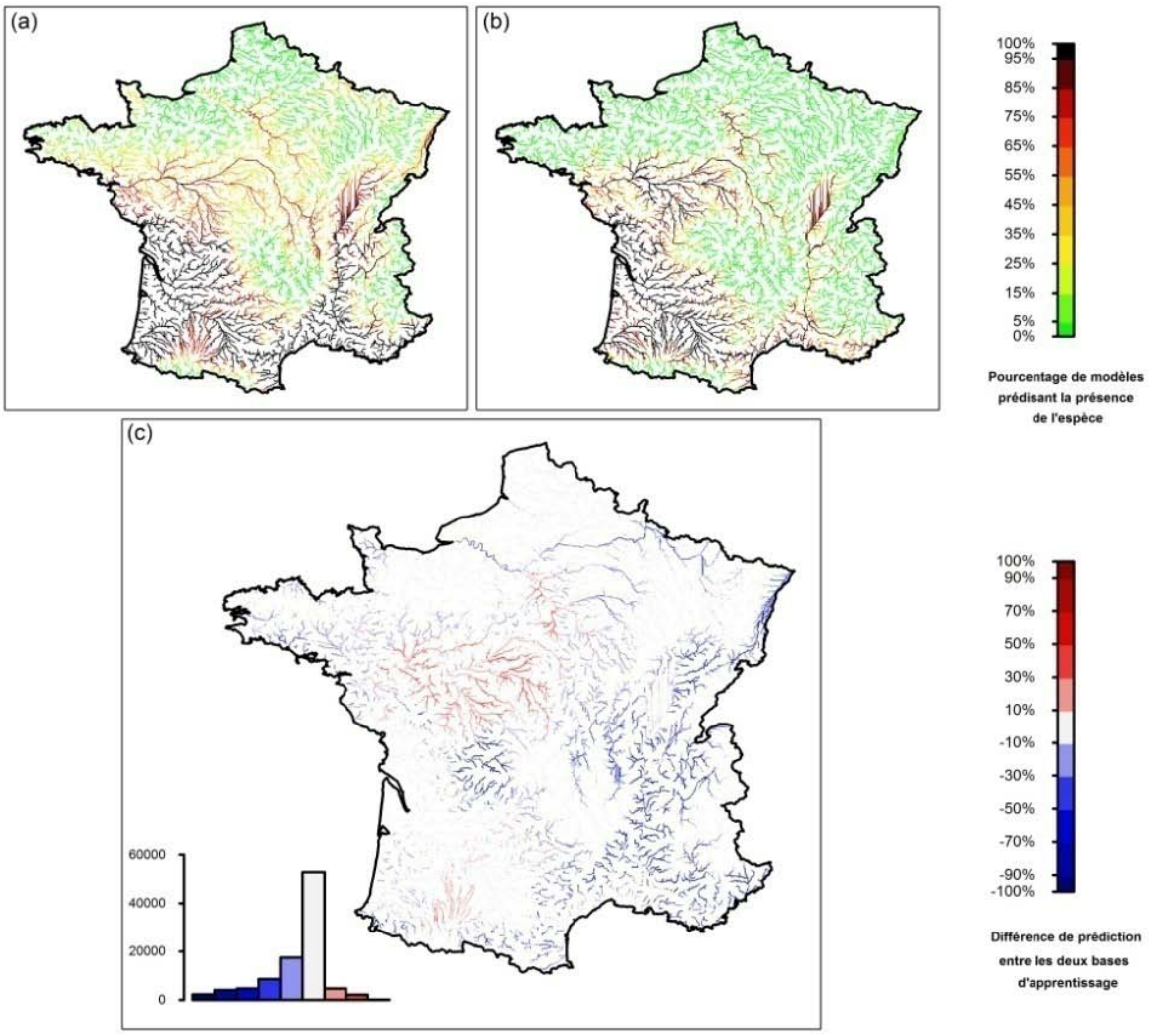


Figure 21 : Influence de la base d'apprentissage sur la distribution actuelle de *Micropterus salmoides* en France métropolitaine. (a) Niveau de risque obtenu en utilisant les occurrences américaines ; (b) Niveau de risque obtenu en utilisant les occurrences françaises. Le risque correspond au pourcentage de modèles prédisant l'espèce comme potentiellement présente à un point donné. (c) Différence entre les deux prédictions (en pourcentage des modèles). Les valeurs négatives correspondent aux pixels prédits comme plus à risque en utilisant la base américaine.

En se basant sur les modèles calibrés sur les données d'occurrence américaines (Figure 22), l'espèce devrait dans le futur étendre son aire de distribution à l'ensemble du territoire français puisque plus de 95 % du territoire serait colonisable en 2050 quels que soient le scénario et le modèle. En 2080, cette proportion diminue pour atteindre 85 % du territoire pour le scénario pessimiste pour le modèle HadCM car le bassin méditerranéen, trop chaud, n'est plus colonisable (Figure 24).

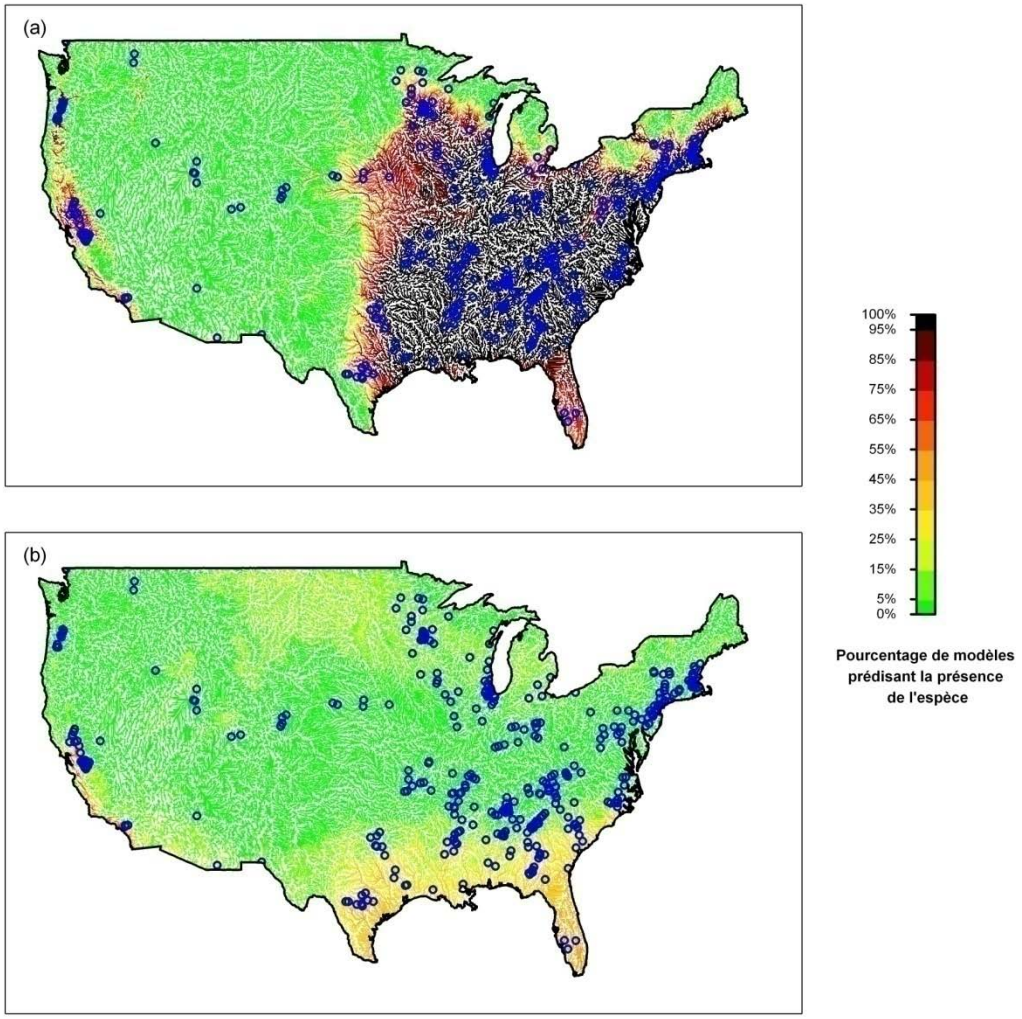


Figure 23 : Influence de la base d'apprentissage sur la distribution actuelle de *Micropterus salmoides* aux USA. (a) Niveau de risque obtenu en utilisant les occurrences américaines ; (b) Niveau de risque obtenu en utilisant les occurrences françaises. Le risque correspond au pourcentage de modèles prédisant l'espèce comme potentiellement présente à un point donné.

b) Ictalurus punctatus

Si pour l'instant cette espèce ne semble susceptible de s'établir que dans le pourtour méditerranéen ainsi que dans les zones aval du bassin du Rhône et les zones médianes du bassin de la Garonne (moins de 10% du territoire présente un risque d'établissement supérieur à 30%), le réchauffement prévu des températures devrait lui permettre d'occuper l'ensemble des bassins Adour-Garonne et Rhône ainsi que les zones médiane et amont des bassins de la Seine et de la Loire, voire plus de 85% du territoire en 2080 dans le scénario le plus

pessimiste, seules les zones méditerranéennes et la pointe de la Bretagne restant épargnées (Figure 25).

c) Clarias gariepinus et Oreochromis mossambicus

Les zones potentielles d'établissement dans les conditions climatiques actuelles (Figures 26, 27) sont situées sur le pourtour méditerranéen et restent à un niveau de risque relativement faible (moins de 50% des modèles prédisant l'établissement) sauf en Corse. Sous l'effet du réchauffement climatique, l'ensemble des zones côtières ainsi que l'aval des grands fleuves seront touchés. Seul, le quart Nord-Est de la France est épargné en 2080 dans le cas du scénario le plus pessimiste (plus de 60% du territoire prédit par plus de 30% des modèles pour *C. gariepinus*, 70% pour *O. mossambicus*).

d) Oreochromis niloticus

Que l'on considère la situation actuelle ou les scénarios futurs, la France ne devrait pas connaitre de conditions environnementales convenables pour cette espèce (Figure 28).

e) Ctenopharyngodon idella

Sous les conditions actuelles, la seule zone d'établissement potentielle est la côte landaise. Dans le futur les zones à risque s'étendent à l'ensemble de la côte atlantique, au cours aval du Rhône et au Jura (Figure 29).

f) Bilan

Malgré des résultats parfois contrastés en fonction des espèces considérées, on note que l'aval des grands fleuves et les zones côtières, en particulier celles des Landes et du pourtour méditerranéen, seraient susceptibles d'accueillir dans le futur la quasi totalité des espèces considérées (Figure 30).

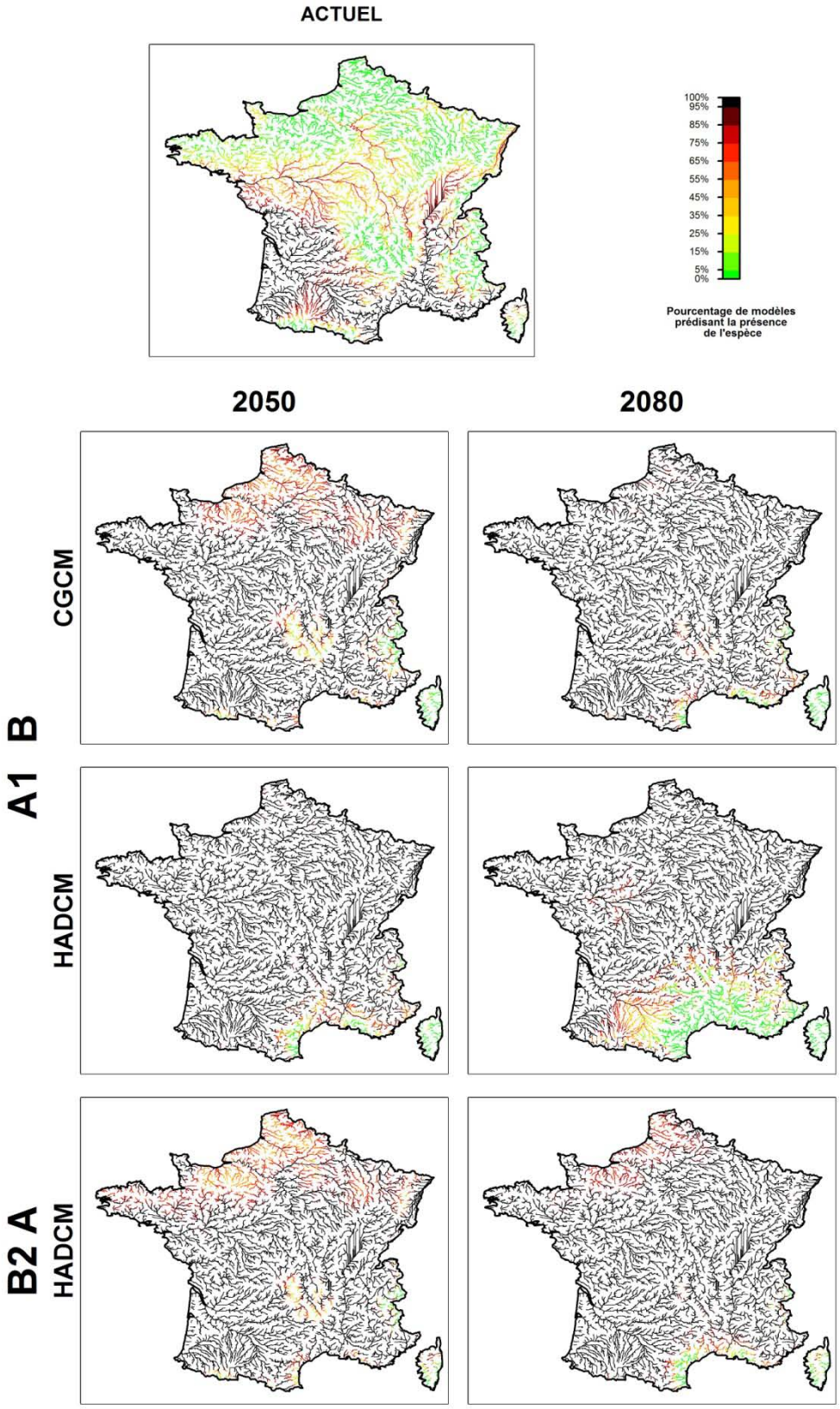


Figure 24 : Risque d'établissement de *Micropterus salmoides* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

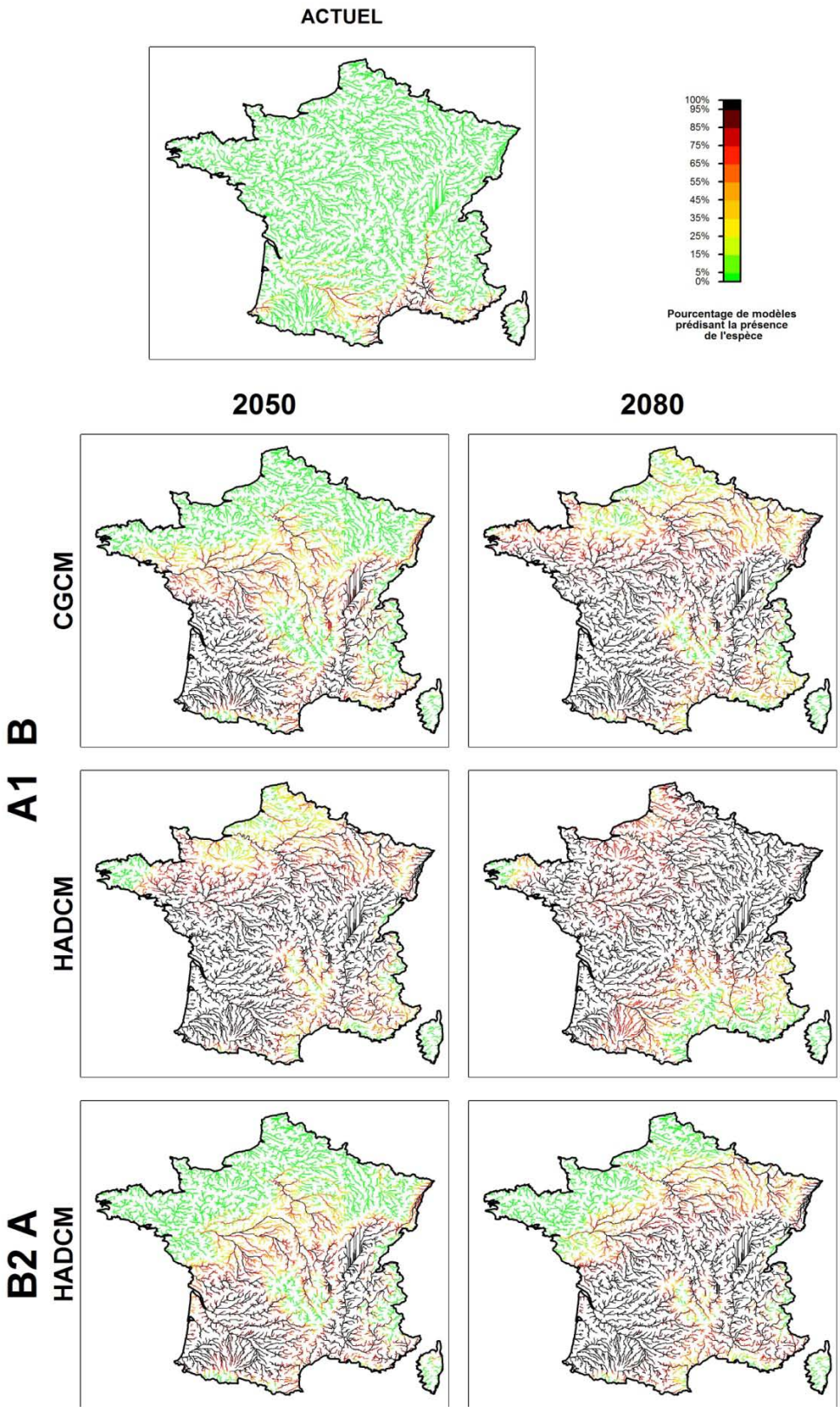


Figure 25 : Risque d'établissement d'*Ictalurus punctatus* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

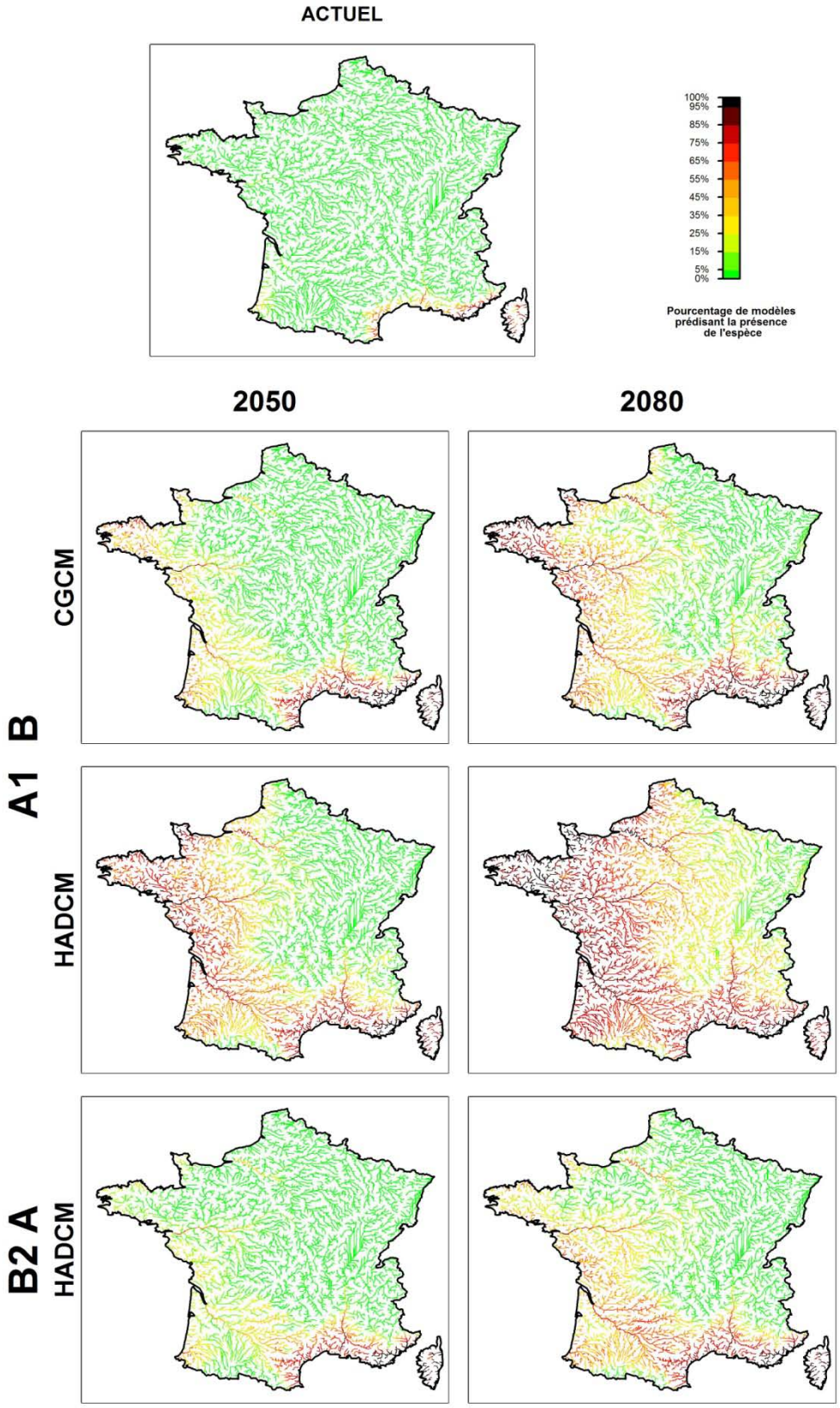


Figure 26 : Risque d'établissement de *Clarias gariepinus* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

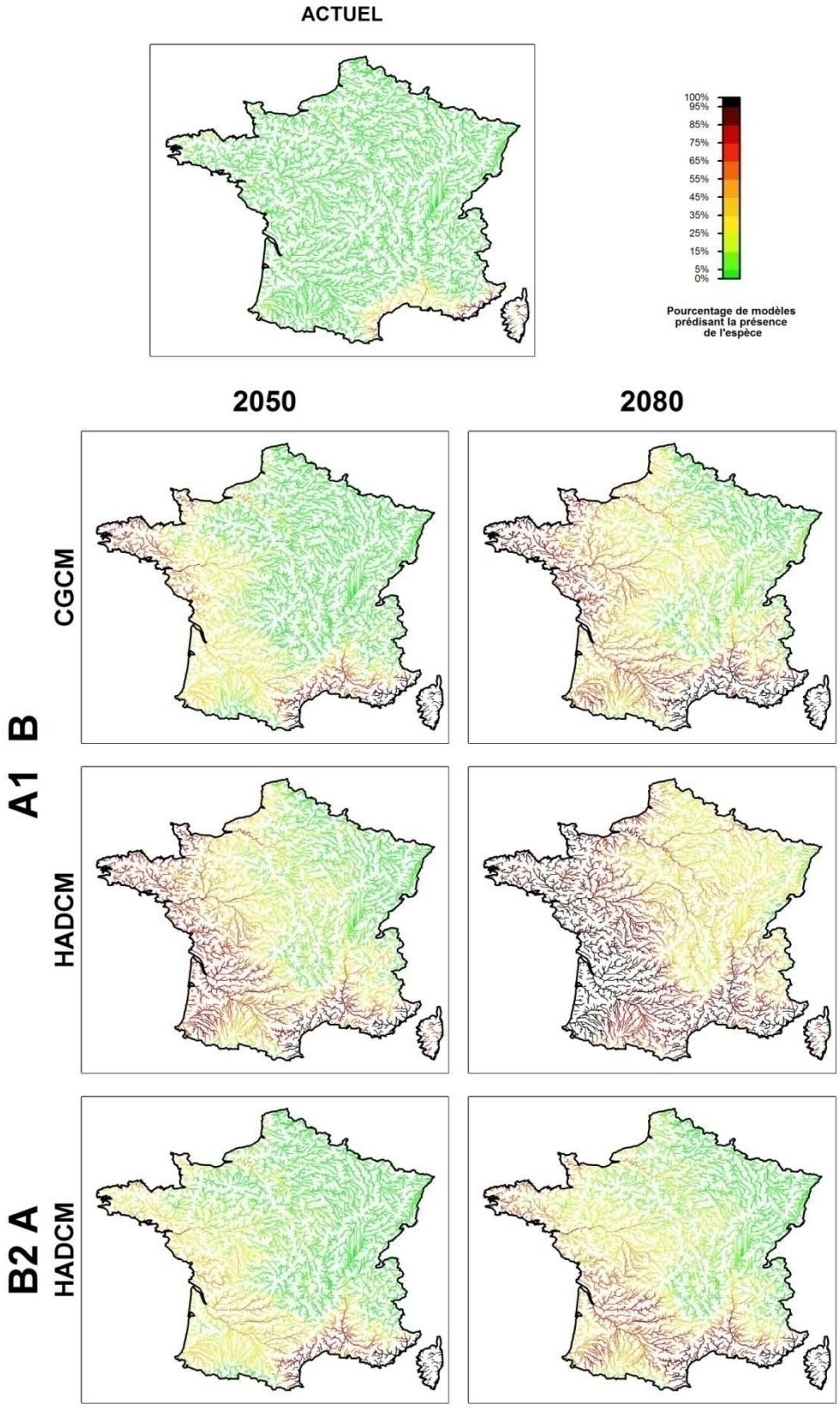


Figure 27 : Risque d'établissement d'*Oreochromis mossambicus* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

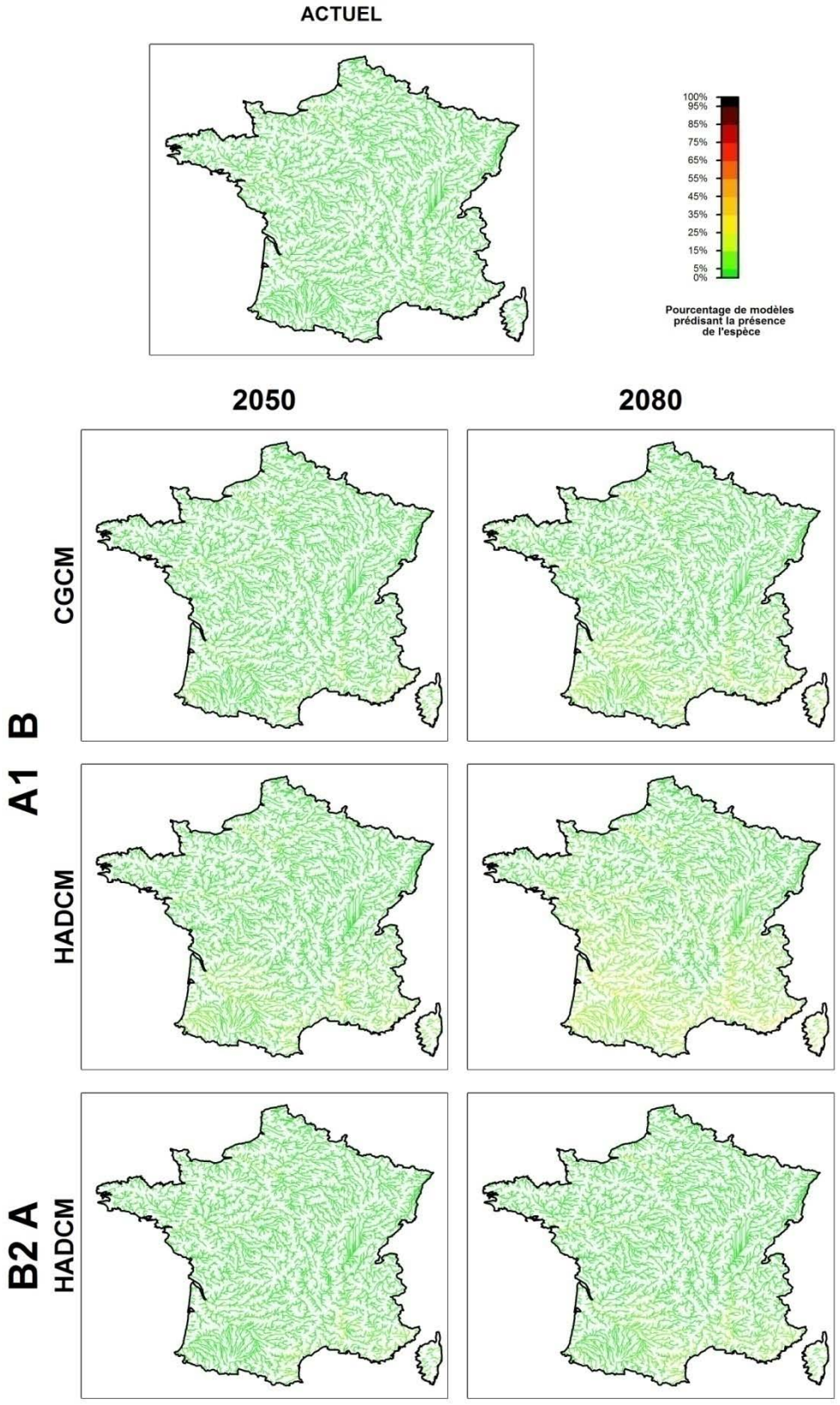


Figure 28 : Risque d'établissement d'*Oreochromis niloticus* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

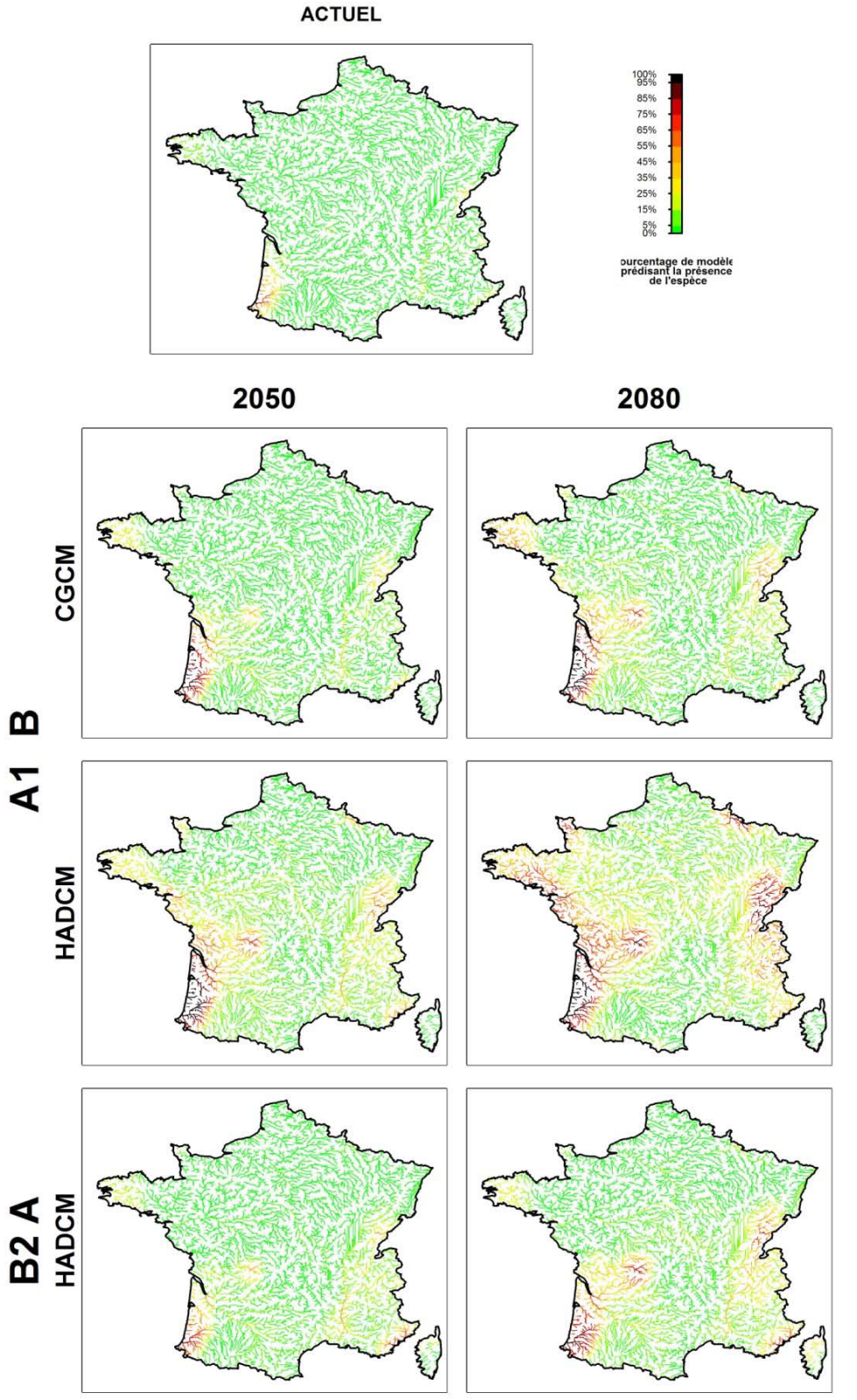


Figure 29 : Risque d'établissement de *Ctenopharyngodon idella* en métropole sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

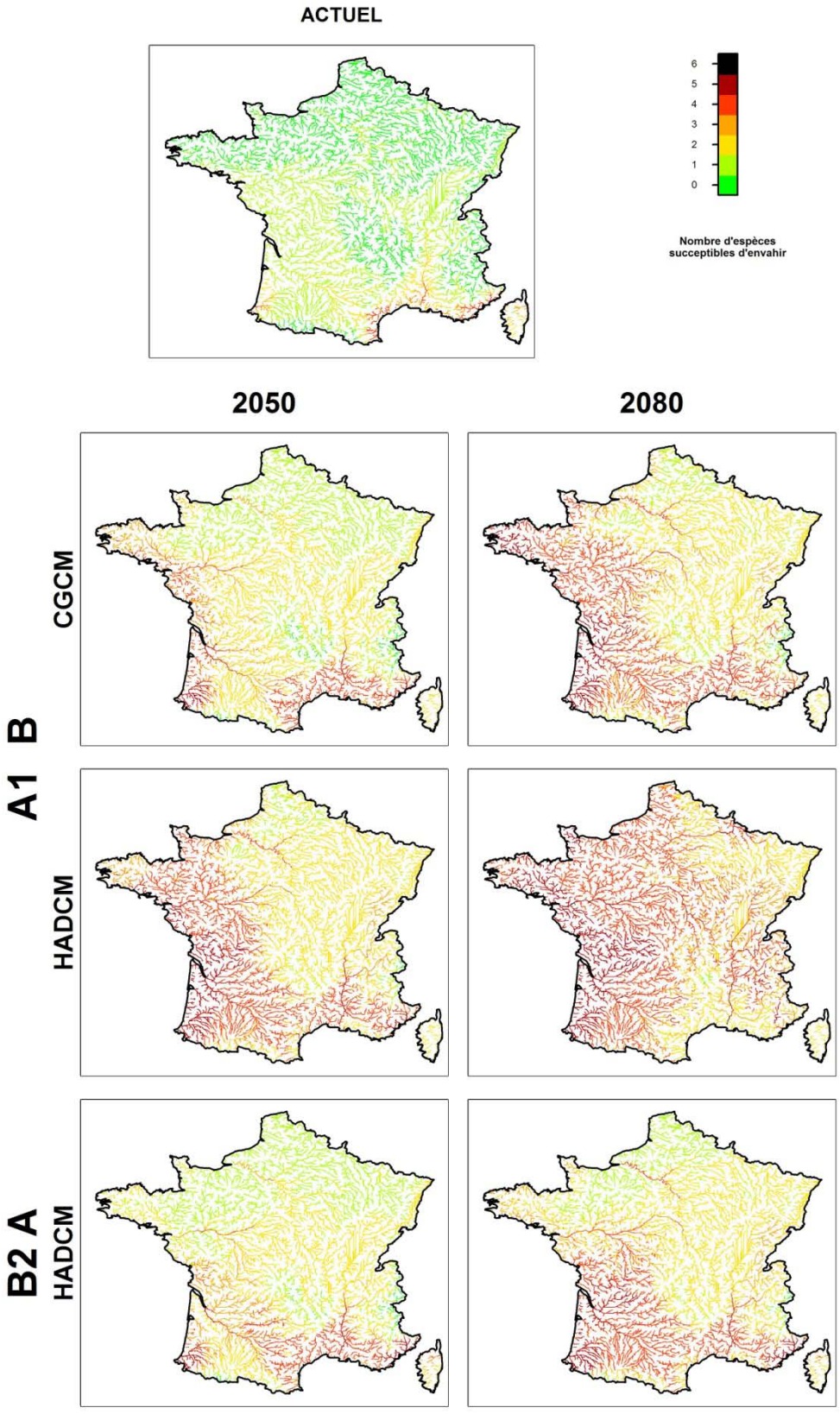


Figure 30 : Nombre d'espèces susceptibles de s'établir (pour chaque pixel, on compte le nombre d'espèces dont la présence est prédite par plus de 30% des modèles).

3. Risques d'établissement dans les DOM

a) Micropterus salmoides et Ictalurus punctatus

Les quatre départements d'Outre Mer, actuellement susceptibles d'héberger l'espèce sur l'ensemble des territoires, devraient voir diminuer le risque d'établissement avec l'élévation des températures sauf dans les cours d'eau guyanais proches de l'océan (Figures 31, 32).

b) Clarias gariepinus

Les conditions climatiques actuelles sont favorables à cette espèce dans les départements insulaires. Avec le réchauffement climatique, c'est l'ensemble des départements d'Outre Mer qui risque d'être touché (Figure 33).

c) Oreochromis mossambicus

Cette espèce ne devrait pas être capable de s'établir dans les territoires d'Outre Mer sauf à la Réunion avec un niveau de risque très élevé aussi bien actuellement que dans le futur (Figure 34).

d) Oreochromis niloticus

Les quatre départements d'Outre Mer sont des zones d'établissement potentiel de l'espèce aussi bien actuellement avec un niveau bas que dans le futur avec un niveau de risque très élevé (Figure 35).

e) Ctenopharyngodon idella

Les quatre départements d'Outre Mer sont des zones d'établissement potentiel de l'espèce aussi bien actuellement avec un niveau élevé que dans le futur avec un niveau de risque plus bas (Figure 36).

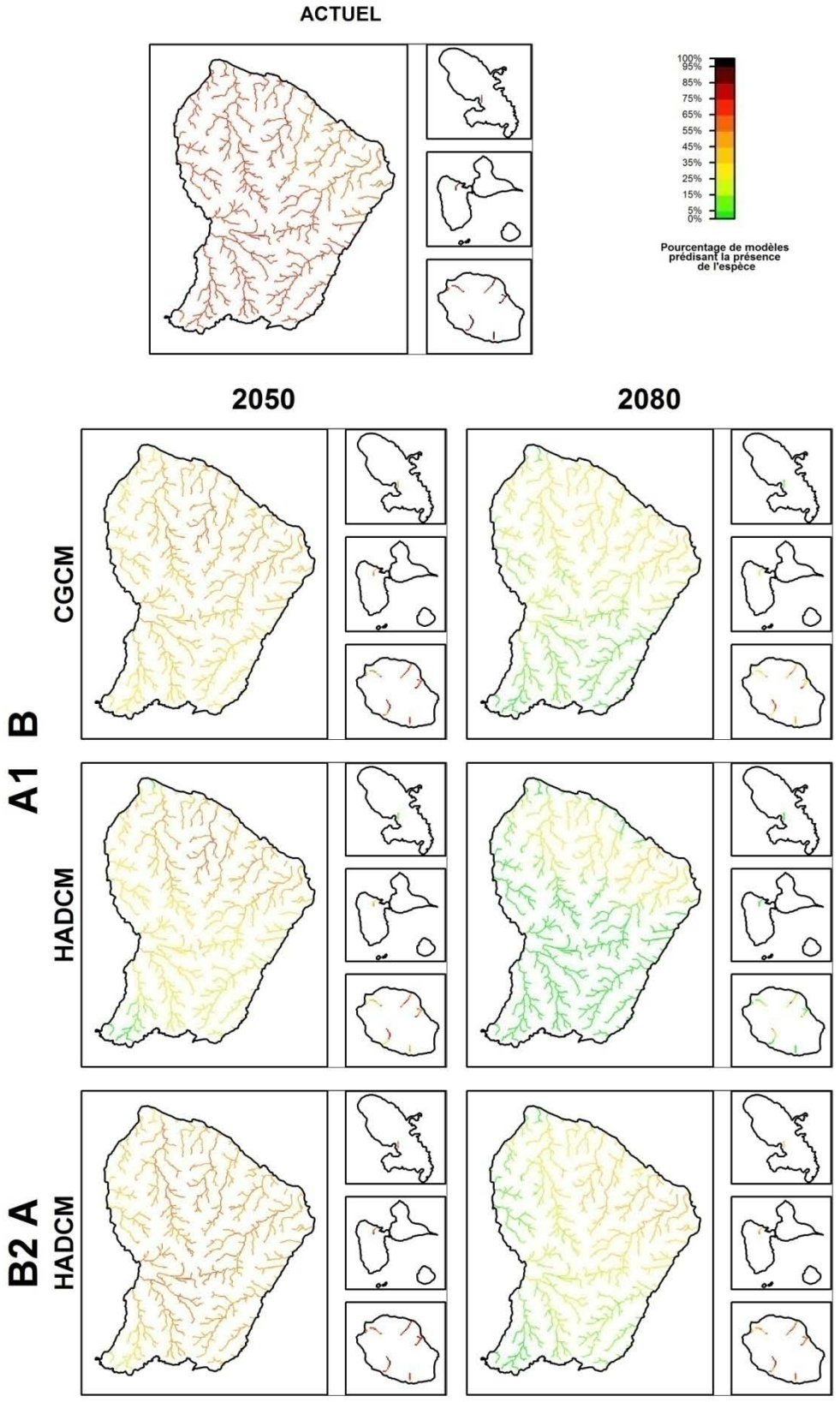


Figure 31 : Risque d'établissement de *Micropterus salmoides* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

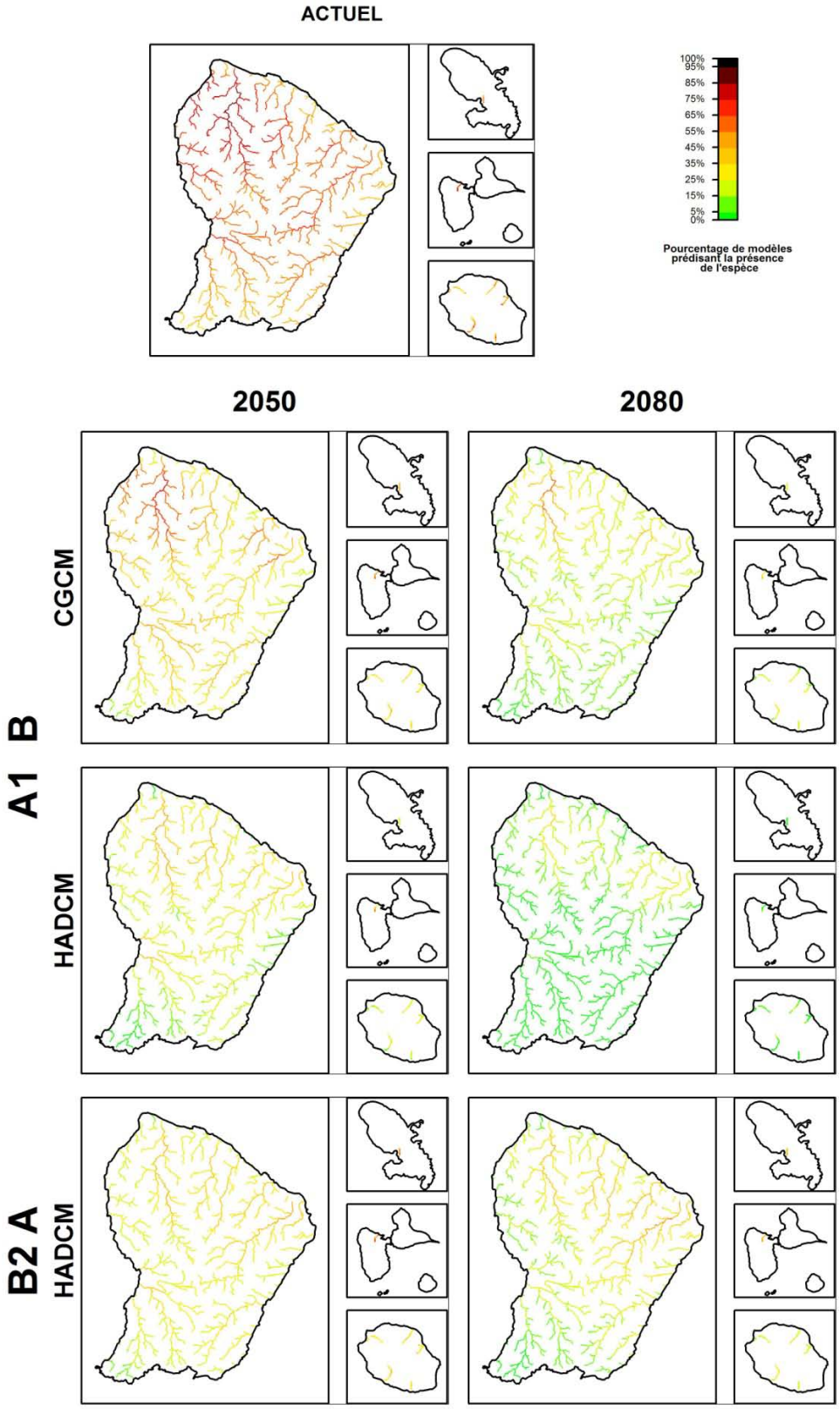


Figure 32 : Risque d'établissement d'*Ictalurus punctatus* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

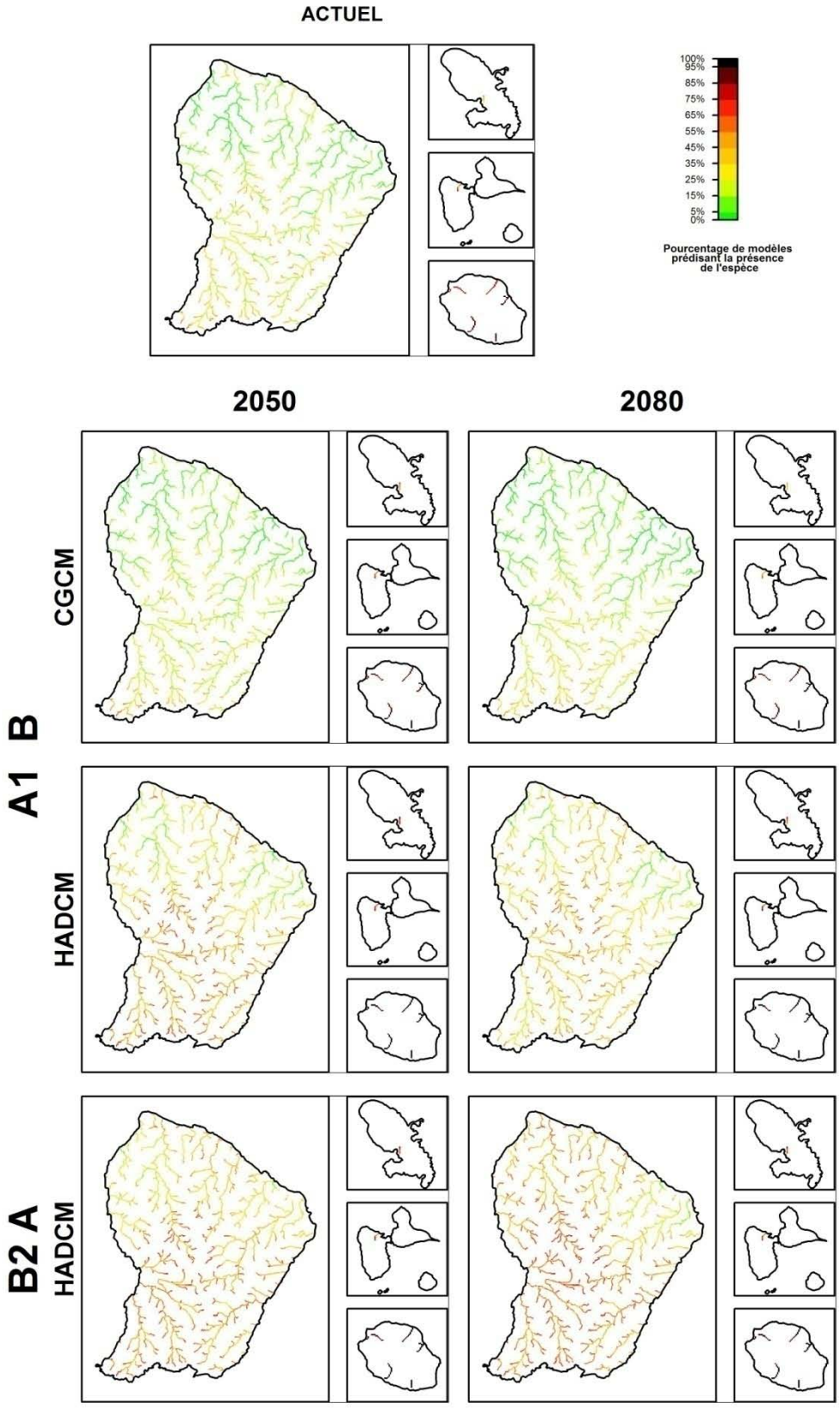


Figure 33 : Risque d'établissement de *Clarias gariepinus* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

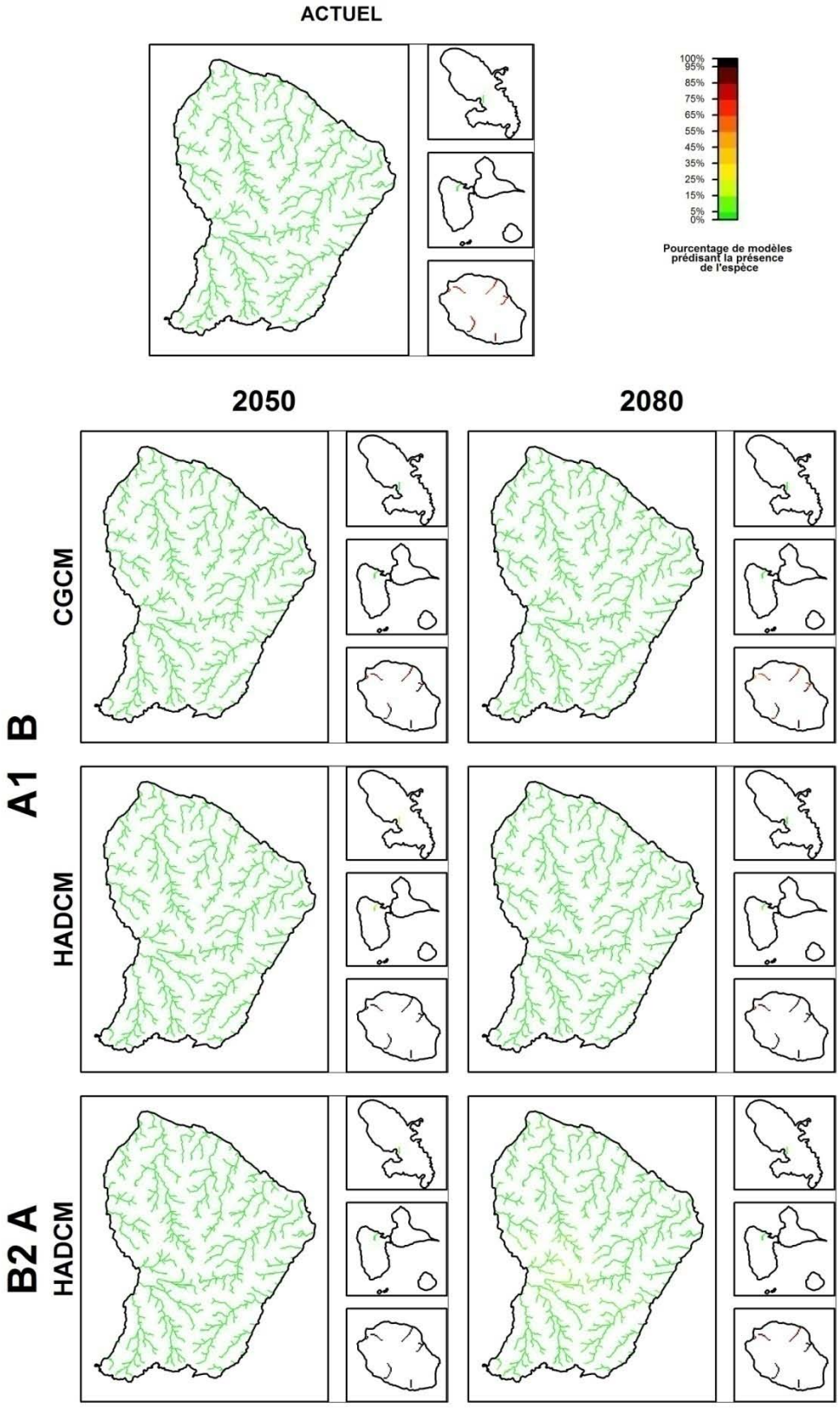


Figure 34 : Risque d'établissement d'*Oreochromis mossambicus* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

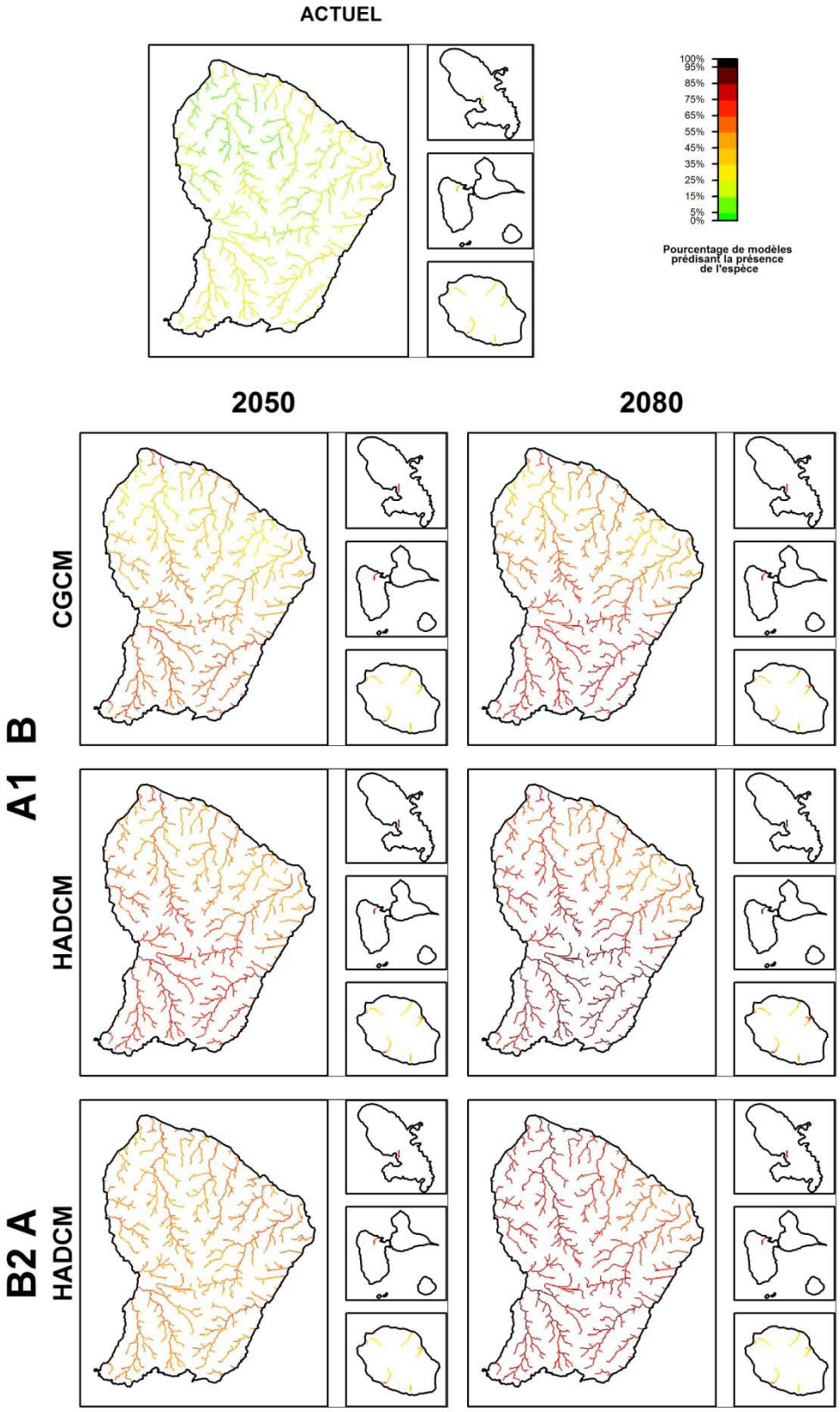


Figure 35 : Risque d'établissement d'*Oreochromis niloticus* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

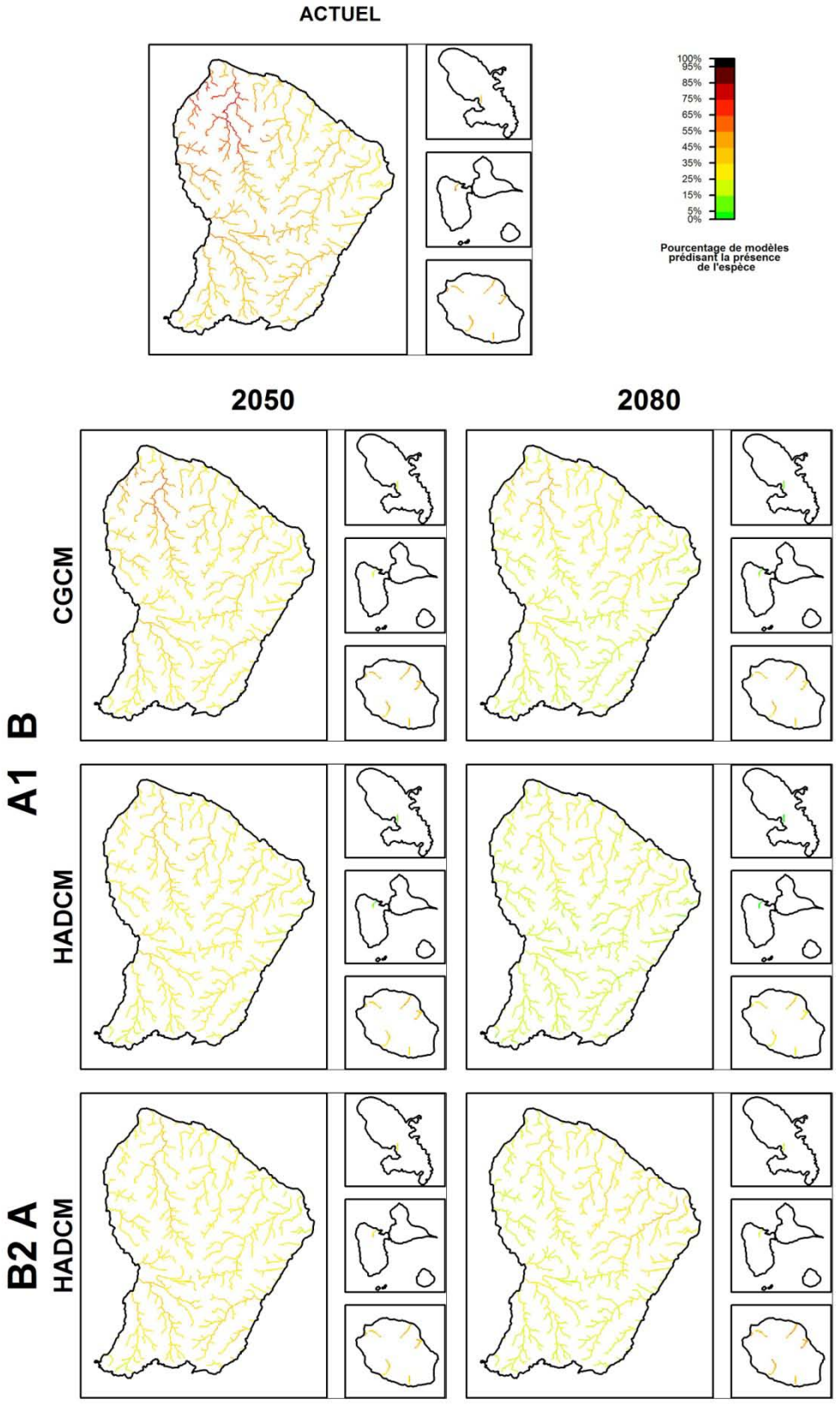


Figure 36 : Risque d'établissement de *Ctenopharyngodon idella* dans les DOM sous les conditions environnementales actuelles et futures. Le risque correspond au pourcentage de modèles prédisant l'espèce comme présente à un point donné.

D. Discussion

La prévention est la méthode de lutte contre les espèces invasives la plus efficace et la moins onéreuse (Moyle and Light 1996; Simberloff and Stiling 1996). Elle passe entre autres par la connaissance de la niche climatique de l'espèce. Les modèles corrélatifs, s'ils permettent d'évaluer la niche climatique de l'espèce à partir de simples données d'occurrence, sont dépendants de la qualité des données (Lobo et al. 2007; Hortal et al. 2008; Lobo et al. 2010) et du choix des variables climatiques utilisées (e.g. Warren and Seifert 2011). La connaissance d'occurrences de *Micropterus salmoides* en France nous a permis de tester la qualité des modèles sur des données totalement indépendantes de celles de la base de calibration. Nous avons pu montrer qu'au moins pour cette espèce, les données d'occurrence disponibles dans l'aire native permettaient de prédire les zones d'établissement de l'espèce déjà observées en France métropolitaine.

La comparaison des distributions obtenues en se basant sur les occurrences soit de l'aire native, soit de l'aire exotique (Hill et al. 2012) confirme que *M. salmoides* n'a pas changé de niche lors de son introduction en France. Les zones françaises prédites avec un risque plus élevé par la base française que par la base américaine s'expliquent par une répartition différente des occurrences le long du gradient amont-aval. En effet, si dans son aire native l'espèce vit préférentiellement dans les zones amont, les occurrences sont beaucoup plus uniformément réparties entre zone amont et zone médiane dans les rivières françaises. Le modèle basé sur les occurrences françaises a donc tendance à prédire des risques plus élevés dans les zones médianes du bassin de la Loire et de la Seine. On note cependant que la niche occupée par l'espèce en France ne correspond probablement qu'à une faible portion de la niche potentielle de l'espèce. En effet, l'utilisation des occurrences françaises ne permet de retrouver qu'une faible fraction de la distribution américaine de l'espèce, les niveaux de

prédiction les plus élevés étant observés hors de la zone de distribution native de l'espèce. Cette très mauvaise qualité des projections sur le continent américain s'explique surtout par le fait que les conditions environnementales rencontrées en France ne sont pas représentatives de l'ensemble des conditions rencontrées dans les sites américains. Les données françaises permettent donc de retrouver la distribution en France mais pas d'extrapoler à des zones dont les conditions environnementales sont très différentes.

Pour les autres espèces, il n'existe malheureusement pas d'autre possibilité pour tester la pertinence des modèles que d'évaluer la qualité des prédictions sur la base de test géographiquement corrélée à celle d'apprentissage. Or on sait que, suivant les espèces, les modèles basés uniquement sur l'aire native sont plus ou moins efficaces pour prédire la distribution de l'espèce dans l'aire exotique (Beaumont et al. 2009). Nos résultats doivent donc être pris avec précaution. Les modèles (de qualité parfois réduite) prédisent l'absence d'*Oreochromis niloticus* de France métropolitaine aussi bien dans les conditions actuelles que sous l'effet du changement climatique. Cependant la présence de l'espèce a déjà été observée en Italie, non seulement dans des rivières thermales, à la température naturellement élevée (Bianco and Turin 2010), mais aussi dans une lagune (Scordella et al. 2003). Il semble donc envisageable que cette espèce, profitant du réchauffement des eaux, puisse coloniser les rivières françaises les plus chaudes en particulier en 2080 dans le cas du scénario le plus pessimiste. La qualité réduite des modèles et l'absence de prédiction de l'espèce en France dans le futur s'expliquent sans doute en partie par le faible nombre d'occurrences de cette espèce (75) ainsi que par le nombre réduit d'occurrences dans la zone septentrionale de son aire de distribution, en particulier le long de la vallée du Nil.

De même, on note l'absence de prédiction de risque pour *O. mossambicus* en Martinique et Guadeloupe où la présence de cette espèce est pourtant avérée (source : ISSG (IUCN)) mais où son établissement est prouvé dans les lacs et retenues mais pas dans les cours d'eau. Les

modèles prédisent l'absence de l'espèce à cause d'une pluviométrie plus abondante que dans l'aire native, ce qui semble cohérent d'un point de vue écologique, car le régime torrentiel des cours d'eau est considéré comme un rempart efficace contre l'établissement des Tilapia (Monti et al. 2010) Si l'espèce se reproduit effectivement dans les Dom, l'amélioration de la qualité des prédictions passerait par l'utilisation de données d'occurrences exotiques les plus nombreuses possibles afin de décrire au mieux l'ensemble de la niche, les distributions potentielles prédites par nos modèles ne correspondant probablement qu'à une partie de la niche climatique. Une sélection des variables climatiques espèce par espèce en fonction des caractéristiques des différentes espèces pourrait également permettre d'affiner les prédictions.

Les résultats concernant les risques d'établissement sont à tempérer par des données écologiques dans le cas de *C. idella*. La reproduction de cette espèce nécessite en effet des conditions hydrologiques particulières : débits importants associés à des températures élevées en période de frai dans de longs segments du cours d'eau (Cudmore and Mandrak 2004). L'établissement de cette espèce semble donc peu probable sauf éventuellement dans les cours inférieurs du Rhône et de la Garonne.

Au contraire, une espèce comme *M. salmoides*, déjà établie en de nombreux points du territoire, ne devrait pas avoir de mal à envahir l'ensemble du territoire dans les décennies à venir, les conditions environnementales lui étant favorables presque partout. Cette espèce prédatrice qui peut se révéler très nuisible pour les écosystèmes receveurs (Shelton et al. 2008; Weyl et al. 2010) a en plus une dispersion facilitée par l'homme de par son intérêt pour la pêche sportive. On peut craindre qu'elle atteigne rapidement l'ensemble des zones favorables à sa reproduction. Il semble donc nécessaire de mettre en place des programmes d'étude pour évaluer l'impact de cette espèce sur les milieux receveurs ainsi que les mesures à mettre en place pour limiter sa dispersion et sa prolifération.

Les trois autres espèces considérées (*O. mossambicus, I. punctatus* et *C. gariepinus*) sont susceptibles de s'installer sur une large portion du territoire métropolitain, en particulier la moitié sud et la façade Atlantique. L'introduction de ces espèces pour l'aquaculture est donc à déconseiller en vertu du principe de précaution, de même que celle d'*O. niloticus* sauf si des études complémentaires confirment l'incapacité de cette espèce à s'établir sur le territoire.

Cette étude a mis en évidence que :

- **les modèles de distribution permettent de prédire la majorité des occurrences actuelles de *M. salmoides*, montrant ainsi leur utilité dans le cadre de la lutte contre les espèces invasives ;**
- **les six espèces étudiées sont susceptibles de s'établir en France (métropolitaine ou DOM) dans les conditions actuelles ou sous l'effet du changement climatique. Leur introduction dans le cadre de l'aquaculture est donc à déconseiller sans étude complémentaire ;**
- ***M. salmoides*, déjà bien présent sur le territoire et pouvant potentiellement coloniser l'ensemble de la France dans les conditions futures, devrait faire l'objet de mesures de contrôle.**

Partie III :
La méthode itérative

A. Problématique

Le changement de niche observé chez certaines espèces invasives limite la qualité des modèles de distribution corrélatifs. Si on n'utilise que les données d'occurrence de l'aire native dans la construction des modèles, on ignore une partie des conditions environnementales qui conviennent à l'espèce. L'utilisation de l'ensemble des données disponibles en incluant l'aire invasive viole quant à elle l'hypothèse fondamentale des modèles corrélatifs supposant que les espèces sont à l'équilibre avec leur environnement, c'est à dire que l'ensemble de la niche potentielle est occupée. En effet, dans le cas d'introductions récentes, l'espèce n'a eu le temps de coloniser qu'une faible partie de sa niche potentielle. Un grand nombre des absences observées dans l'aire exotique sont des absences contingentes (et non environnementales). Les modèles corrélatifs construits en utilisant ces données tendent alors à sous-estimer la distribution potentielle (e.g. Beaumont et al. 2009).

Si les modèles presence-only permettent de contourner au moins partiellement ce problème, ils conduisent souvent à une surestimation de l'aire potentielle car ils ne prennent pas en compte les vraies absences qui contiennent une part non négligeable de l'information sur la distribution de l'espèce. De plus, ces modèles, en utilisant un background (l'ensemble des conditions environnementales présentes dans la zone d'étude), sont eux aussi influencés par les données d' « absence » (Sexton et al. 2002; Phillips et al. 2009) . L'utilisation de modèles de présence/absence ne prenant en compte que les points d'absence de l'aire native peut difficilement être envisagée car si les données d'absence ne couvrent pas une gamme suffisante de conditions environnementales, on se retrouve confronté à des problèmes d'extrapolation. La tendance actuelle est d'utiliser toutes les données de présence disponibles aussi bien dans l'aire native que dans l'aire exotique (Beaumont et al. 2009; Capinha et al. 2011; Jiménez-Valverde et al. 2011) et des pseudo-absences (des sites choisis aléatoirement dans la zone d'étude et qu'on considère comme des absences puisque la présence de l'espèce

n'y a pas été vérifiée). L'utilisation de pseudo-absences est particulièrement utile quand une partie des données provient de sources telles que des échantillons de muséum et ne contient donc aucune information d'absence.

Le mode de sélection des pseudo-absences a fait l'objet de nombreuses études. Il a été proposé de les sélectionner en utilisant des modèles presence-only (Engler et al. 2004), mais cette méthode peut conduire à la sélection de nombreuses absences peu informatives, leurs conditions environnementales étant très éloignées de celles de la niche (Lobo et al. 2010). La localisation des pseudo-absences est cependant moins importante que leur nombre et les biais dans l'échantillonnage des présences (Lobo et al. 2010) et l'utilisation de pseudo-absences aléatoires sur l'ensemble de la zone de projection semble le meilleur choix pour la modélisation des espèces invasives (Capinha et al. 2011).

L'inconvénient majeur des deux méthodes précédentes (presence-only et pseudo-absences) est de ne pas utiliser les données d'absence contenues dans la base de données, dont certaines sont informatives. L'idéal serait de pouvoir identifier et éliminer les absences non environnementales. Cela est envisageable pour les absences méthodologiques (l'espèce n'a pas été observée alors qu'elle était présente) en prenant en compte la détectabilité de l'espèce. Des modèles de probabilité binomiale ont été utilisés avec succès pour étudier le changement de gamme d'altitude des espèces sous l'effet du changement climatique (Moritz et al. 2008; Rowe et al. 2010) et pour modéliser la distribution des espèces en utilisant des modèles d'« occupation de sites ». Malheureusement ce type de méthodes nécessite des données d'abondance ou un échantillonnage répété des mêmes sites. Elle n'est donc utilisable que sur un nombre réduit d'espèces et ne résout pas le problème des absences contingentes (l'espèce n'est pas présente car elle n'a pas eu l'occasion de s'installer suite à des contraintes de dispersion ou des interactions biotiques).

B. La méthode itérative

Nous avons cherché à construire une méthode basée sur les modèles de présence/absence qui puisse s'appliquer à un large éventail d'espèces. Pour cela nous nous sommes inspirés de méthodes de modélisation couramment utilisées en physique et basées sur un processus itératif. Les données issues d'une simulation sont considérées comme représentant le nouvel état du système après une petite période d'évolution et utilisées comme données initiales du modèle pour une nouvelle simulation. Le processus s'arrête quand les résultats ne différent plus significativement d'une simulation à la suivante. On peut alors estimer que le système a atteint un état stable. Nous avons adapté cette méthode pour mimer un processus d'invasion et ainsi atteindre un état virtuel d'équilibre de l'espèce concernée avec son environnement. L'état initial du système correspond à l'occupation observée des sites. La première étape de modélisation (un modèle corrélatif classique utilisant une méthode d'ensemble) permet de prédire tous les sites colonisables à partir des sites d'occupation initiaux. Cette distribution prédite est ensuite vue comme l'ensemble des sites sources de la colonisation potentielle. Les données de présences étant beaucoup plus fiables que celles d'absences nous avons souhaité leur donner un poids supplémentaire. Pour cela, à la fin de chaque étape du processus itératif, nous avons considéré que la distribution potentielle était constituée non seulement de la distribution prédite par le modèle mais aussi des sites où l'espèce avait été observée mais qui étaient omis par le modèle (fausses absences). En d'autres termes, le vecteur des occurrences utilisé à la nième itération correspond au vecteur des occurrences observées dans lequel on a remplacé les absences (0) par des présences (1) si le modèle construit à l'itération n-1 prédisait une présence. Le processus est répété jusqu'à ce que les prédictions se stabilisent. Pour évaluer cette stabilisation, nous avons du prendre en compte les variabilités des prédictions dues aux méthodes statistiques utilisées. Le système est dit stabilisé dès que la variabilité des prédictions est inférieure ou égale à ce qui est observé en une seule étape de

modélisation à partir de la même base d'apprentissage en répétant plusieurs fois les simulations.

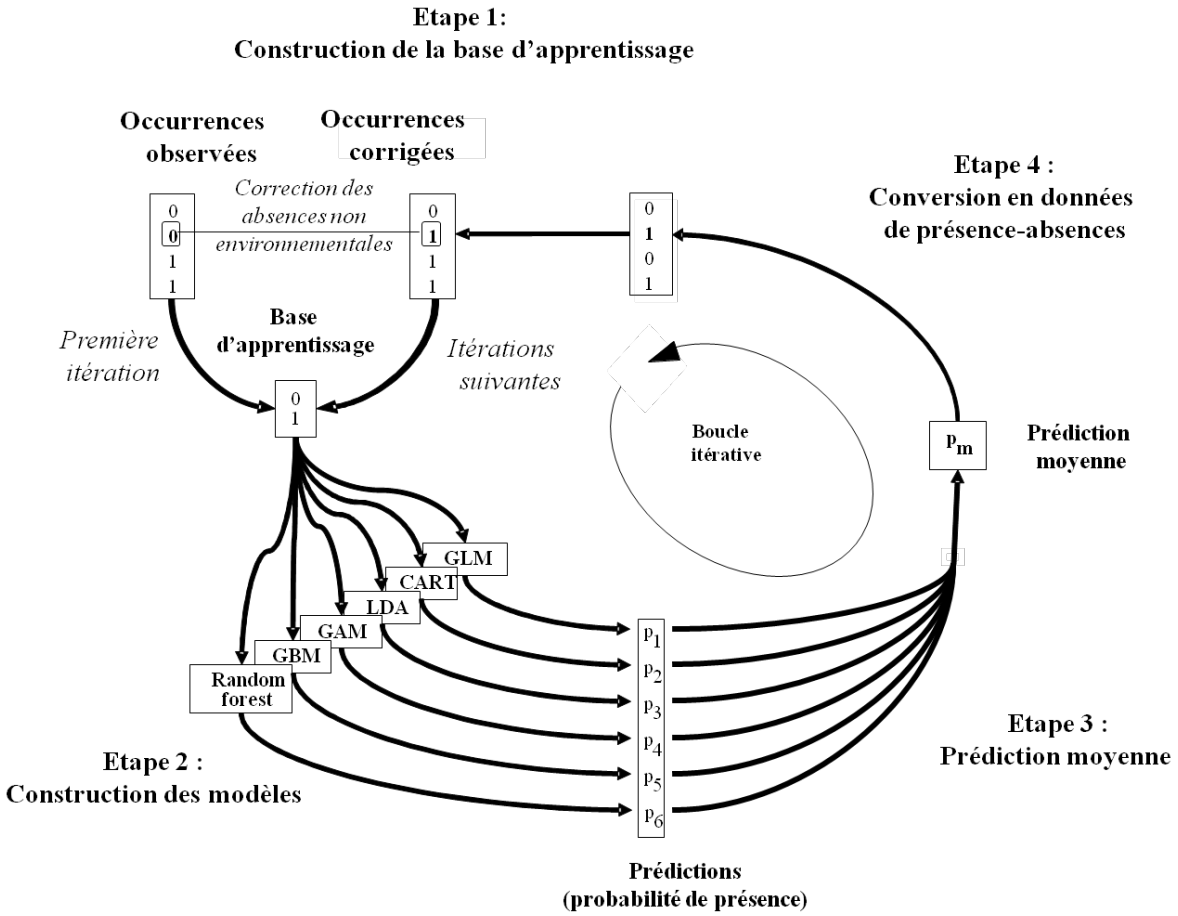


Figure 37 : Le principe de la méthode itérative.

C. Les espèces et les données environnementales

1. Pour l'étude de la méthode itérative basée sur des espèces virtuelles (M3)

Cette méthode itérative ayant été construite dans le but de corriger les absences non environnementales, il était souhaitable de partir d'une base dans laquelle on connaissait précisément la nature (présence ou absence) de chaque site. Cette contrainte est difficilement remplie dans les bases de données sur des espèces réelles, à plus forte raison si elles sont mobiles, et donc généralement de détectabilité inférieure à 1. Elle reste peu réaliste même

pour de nombreuses espèces non-mobiles, une absence locale pouvant être liée à des interactions biotiques ou à des évènements aléatoires (broutage par un herbivore…). Pour éviter ce biais, nous avons testé la méthode itérative sur des espèces virtuelles construites sur le même principe que celles utilisées dans l'étude (M1). Des absences ont été introduites dans la base d'apprentissage de deux façons différentes : aléatoirement et en respectant un gradient de détectabilité. En effet, dans le cas des espèces réelles, la densité de l'espèce a tendance à être plus importante à l'optimum des conditions environnementales (Brown 1984), ce qui augmente la détectabilité de l'espèce. Les absences méthodologiques sont donc généralement plus nombreuses en bordure de la niche environnementale.

2. Pour l'étude de la méthode itérative basée sur des espèces réelles de poissons (M4)

Nous avons ensuite testé la méthode itérative sur des espèces réelles. Pour cela nous avons utilisé les données fournies par l'ONEMA. Ces données couvrent plus de 20 ans de campagnes de suivi par pêche électrique des communautés piscicoles françaises. Les 1110 stations d'échantillonnage utilisées dans notre étude sont réparties sur l'ensemble du territoire français et sont représentatives des différents types de cours d'eau.

Pour mesurer l'efficacité de la méthode itérative à combler les absences non environnementales, nous devions disposer d'une base de test contenant un nombre le plus réduit possibles d'absences méthodologiques. Pour cela, nous avons tout d'abord sélectionné les sites échantillonnés au moins 15 fois. Pour chacun de ces sites, nous avons ensuite calculé une courbe de saturation moyenne basée sur 100 classements chronologiques aléatoires des campagnes. Nous avons conservé les 191 sites pour lesquels moins d'une espèce apparaissait sur la courbe moyenne lors des 5 dernières pêches. Une espèce était considérée comme présente sur ce site si elle y avait été observée lors d'au moins une des campagnes

d'échantillonnage. Les 919 sites restants ont été utilisés pour calibrer les modèles. Pour ces sites nous avons utilisé les données d'occurrence des espèces obtenues lors d'une seule campagne récente d'échantillonnage.

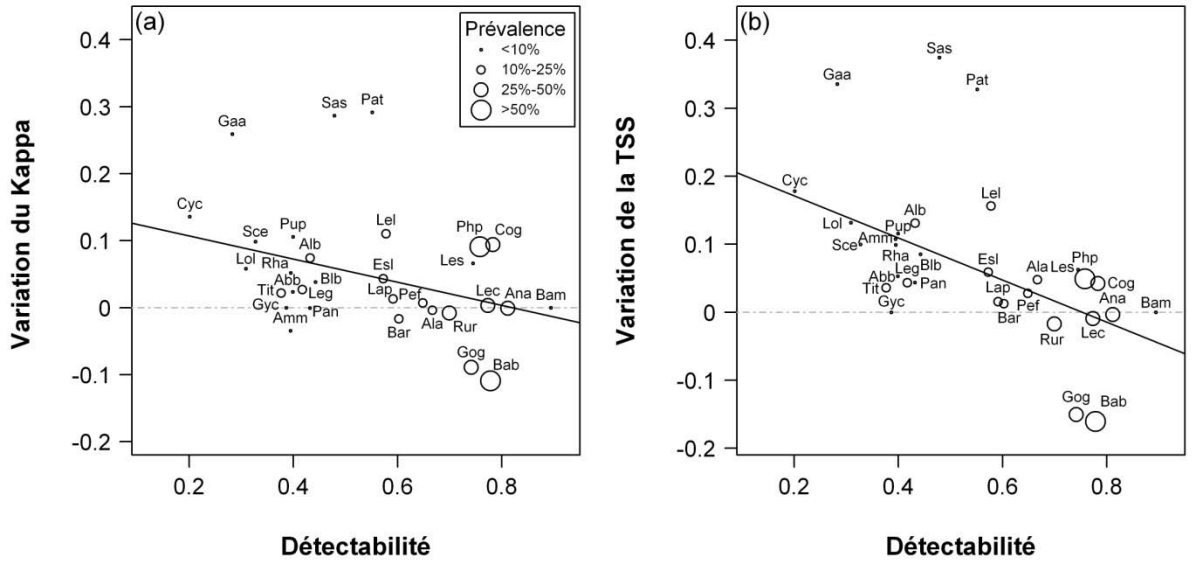


Figure 38 : Variation de la qualité des modèles entre la méthode d'ensemble classique (EM) et la méthode itérative (IEM) en fonction de la détectabilité pour les 31 espèces de poissons. La qualité des modèles est calculée sur les 191 sites bien échantillonnés. a) Kappa, b) TSS. La taille des points est proportionnelle à la prévalence des espèces dans la base d'apprentissage. Les droites représentent la relation entre la détectabilité et la variation de qualité des modèles (Kappa : $p<0.05$; TSS : $p<0.01$). **Codes espèces :** *Abramis brama* (Abb) ; *Alburnus alburnus* Ala), *Alburnoides bipunctatus* (Alb) ; *Ameiurus melas* (Amm) ; *Anguilla anguilla* (Ana) ; *Barbatula barbatula* (Bab) ; *Barbus meridionalis* (Bam) ; *Barbus barbus* (Bar) ;*Blicca bjoerkna* (Blb) ; *Cottus gobio* (Cog) ; *Cyprinus carpio* (Cyc) ; *Esox lucius* (Esl) ; *Gasterosteus aculeatus* (Gaa) ; *Gobio gobio* (Gog) ; *Gymnocephalus cernuus* (Gyc) ; *Lampetra planeri* (Lap) ; *Lepomis gibbosus* (Leg) ; *Leuciscus leuciscus* (Lel) ; *Leuciscus souffia* (Les) ; *Lota lota* (Lol) ; *Parachondrostoma nasus* (Pan) ; *Parachondrostoma toxostoma* (Pat) ; *Perca fluviatilis* (Pef) ; *Phoxinus phoxinus* (Php) ; *Pungitius pungitius* (Pup) ; *Rhodeus amarus* (Rha) ; *Rutilus rutilus* (Rur) ; *Salmo salar* (Sas) ; *Scardinius erythrophtalmus* (Sce) ; *Squalius cephalus* (Sqc) ; *Tinca tinca* (Tit).

Les modèles (EM et IEM) ont été construits en suivant la même méthodologie que dans le cas des espèces virtuelles. Seules les variables environnementales et climatiques diffèrent. Nous avons utilisé deux variables topographiques permettant de résumer la position du site dans le cours d'eau et la vélocité locale (Buisson et al. 2008), et trois variables climatiques : la température annuelle moyenne, l'amplitude thermique entre le mois le plus chaud et le mois le plus froid et la précipitation annuelle moyenne (Buisson et al. 2008). Ils nous ont servi à modéliser la distribution de 31 espèces de poissons dont la prévalence dans la base de calibration était supérieure à 2.5%.

D. Discussion

En utilisant les espèces virtuelles, nous avons pu montrer que la méthode itérative améliorait la qualité des prédictions pour tous les indices de qualité utilisés sauf l'AUC (qui diminue très légèrement) dès que le pourcentage de fausses absences était supérieur à 30%. Pour des taux de fausses absences faibles, les résultats sont plus mitigés puisqu'on observe soit une augmentation, soit une diminution, soit une absence de variation des indices suivant l'espèce et l'indice utilisés. Ces résultats ont été confirmés en utilisant des espèces réelles, puisque la qualité des modèles était améliorée pour les espèces à faible détectabilité, dont la base d'apprentissage contenait un grand nombre d'absences, les espèces les mieux détectées (avec un faible taux de fausses absences) ne profitant pas des itérations (Figure 38). La méthode itérative semble donc plus adaptée que les méthodes classiques pour prédire la distribution potentielle des espèces dont la base d'occurrences contient de nombreuses absences non environnementales, qu'elles soient méthodologiques ou contingentes.

La capacité de la méthode itérative à combler les fausses absences permet d'augmenter le compromis entre prédictions provenant de modèles calibrés sur des bases d'apprentissage différentes (Figure 39), soit avec une répartition différente des fausses absences et donc de limiter les biais liés à l'échantillonnage qui peuvent fortement impacter la qualité des modèles (Lobo et al. 2007; Hortal et al. 2008; Stankowski and Parker 2011). La méthode itérative présente aussi l'avantage d'augmenter le compromis entre les prédictions des différents modèles statistiques. Cela est d'autant plus intéressant que la méthode statistique est une des sources principales d'incertitude dans le cadre des prédictions sous l'effet du changement climatique (Buisson et al. 2010).

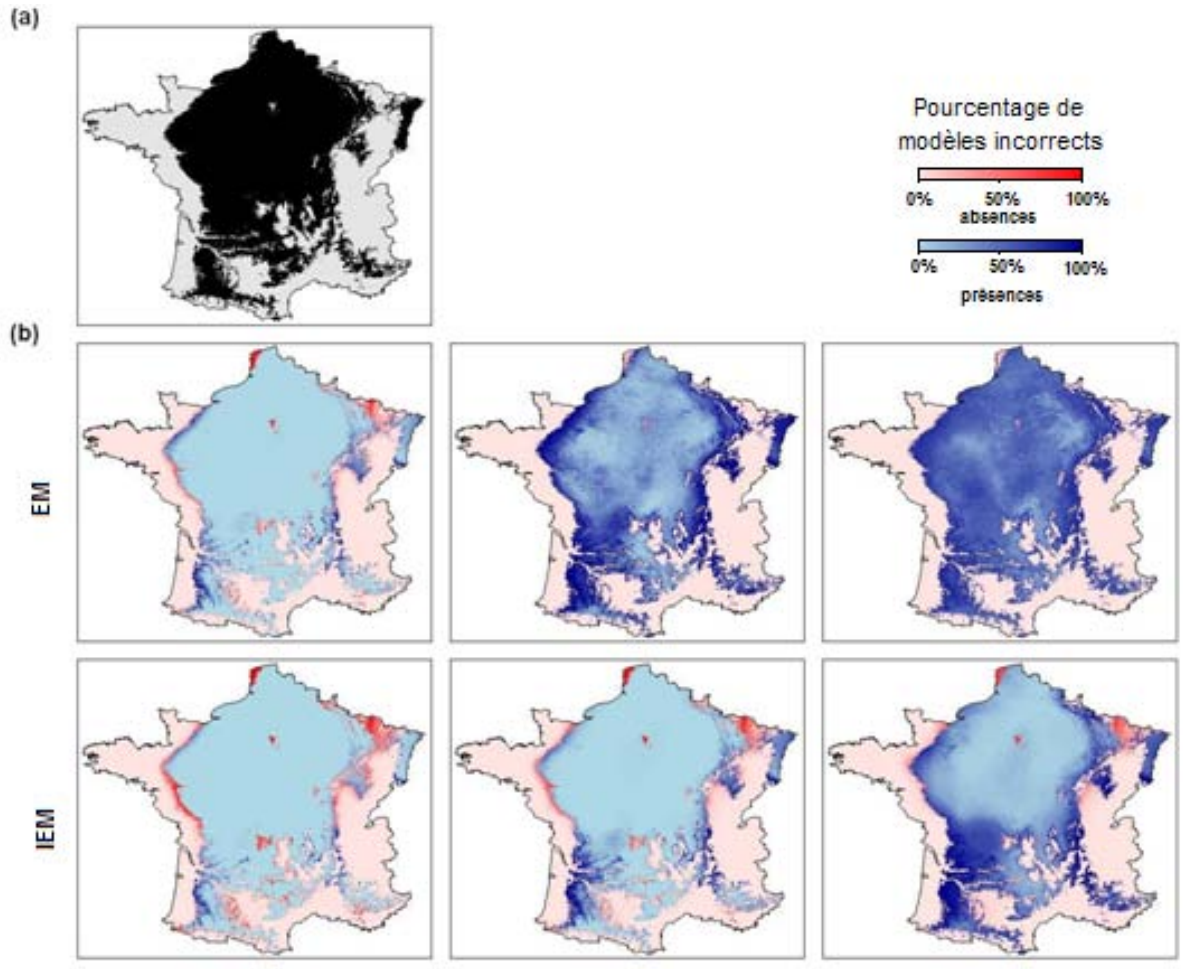


Figure 39 : Comparaison des prédictions de distribution obtenues avec la méthode d'ensemble classique (EM) et la méthode itérative (IEM). a) Niche observée de l'espèce virtuelle de prévalence 60%. b) Prédictions obtenues avec la méthode d'ensemble classique (haut) et la méthode itérative (bas) pour 15% (gauche), 45% (centre) et 75% (droite) d'absences situées aléatoirement dans la niche. Un pixel est d'autant plus foncé que le nombre de modèles qui prédisent incorrectement ce pixel est élevé.

La méthode itérative semble prometteuse pour élargir le champ d'application des modèles de distribution corrélatifs à un grand nombre d'espèces jusque là difficiles à modéliser. Elle peut s'appliquer aux espèces invasives pour lesquelles de nombreuses absences de la zone exotique sont contingentes (e.g. Václavík and Meentemeyer 2011). Elle devrait également permettre d'améliorer les prédictions de distribution des espèces en danger. En effet, celles-ci ont souvent été extirpées d'une large partie de leur aire de distribution. Les absences non environnementales sont donc ici aussi très nombreuses. La méthode itérative devrait aussi permettre d'améliorer la qualité des méthodes hybrides. En effet, dans ces modèles, certains

paramètres comme les interactions biotiques peuvent être pris en compte deux fois (une fois dans le modèle mécanistique dans les paramètres physiologiques, une fois dans le modèle corrélatif via les absences contingentes). Cet « effet cyclique » , qui peut conduire à une sous estimation de la niche (Gallien et al. 2010), devrait être réduit grâce à la méthode itérative.

Cependant, la méthode itérative reste incapable de « deviner » l'information écologique manquante et ne peut combler les fausses absences que si les présences connues couvrent l'ensemble de l'étendue des conditions environnementales favorables à l'espèce. C'est pourquoi on note que dans le cas des espèces virtuelles, si les absences sont plutôt situées en bordure de niche et en grand nombre, la méthode itérative a du mal à retrouver le bord de la niche. De la même façon, elle n'améliore pas les prédictions des deux espèces de poissons ayant la plus faible prévalence (poisson chat et grémille) et pour lesquels les données de présence ne couvrent qu'une petite partie de la niche.

La méthode itérative nécessite également des études complémentaires avant que son usage soit généralisé. Le choix des variables utilisées dans les modèles, qui est une étape cruciale de la modélisation (Wisz and Guisan 2009; Jiménez-Valverde et al. 2011), est d'autant plus important avec notre méthode que sa nature itérative risque d'augmenter les erreurs liées à un mauvais choix. Une étude de l'impact de l'introduction de variables non pertinentes dans les modèles est donc indispensable. De la même façon, il est important d'évaluer l'effet des fausses présences sur les prédictions, leur occurrence en trop grand nombre dans la base de calibration risquant de conduire à une surestimation de la distribution potentielle de l'espèce. Pour finir, l'impact des paramètres connus pour affecter les prédictions des modèles de niche : prévalence de l'espèce, choix de la valeur seuil, échantillonnage des variables environnementales (Liu et al. 2005; Menke et al. 2009; Williams et al. 2009) doit être évalué.

Les deux études, portant à la fois sur des espèces virtuelles et des espèces réelles, ont mis en évidence que :

- **les fausses absences peuvent dégrader notablement la qualité des prédictions des modèles de distribution corrélatifs ;**
- **les fausses absences sont à l'origine d'une grande variabilité dans les prédictions en fonction du choix de la base d'apprentissage ;**
- **la méthode itérative permet de réduire les erreurs des modèles liées aux absences non environnementales ;**
- **la méthode itérative réduit la variabilité des prédictions liée au choix de la base d'apprentissage ;**
- **la méthode itérative augmente le consensus des prédictions obtenues avec des méthodes statistiques différentes.**

Conclusion et perspectives

A. Conclusion

Les modèles de distribution sont très largement utilisés en écologie en particulier dans le cadre du changement climatique. Cette utilisation rencontre des difficultés d'origine aussi bien écologique que méthodologique. De nombreux paramètres introduisent de l'incertitude dans les prédictions de distribution (échelle, variables prédictives, modèles statistiques, base d'apprentissage, scénarios climatiques, modèles de circulation…). Les incertitudes dues aux modèles statistiques sont partiellement contrôlées par l'utilisation de méthodes d'ensemble (Araújo and New 2007) et nous avons montré que le grain avait un effet relativement réduit sur la qualité des prédictions actuelles. Nous avons par contre mis en évidence une forte variabilité des prédictions en fonction de la base d'apprentissage due aux absences non environnementales et nous avons proposé une méthode pour améliorer la qualité des modèles dans le cas où ces absences sont nombreuses. La méthode itérative devrait en particulier permettre d'améliorer les prédictions de distribution des espèces invasives, qui sont une menace majeure pour les écosystèmes et en particulier les milieux aquatiques d'eau douce (Cucherousset and Olden 2011).

Des interactions croissantes entre les réseaux d'observation, qui collectent sur le terrain les données d'occurrence des espèces, et les modélisateurs, qui intègrent ces données dans des modèles d'efficacité croissante, permettant de cibler les zones les plus à risque et de mesurer l'efficacité des mesures de contrôle, devraient permettre de limiter l'impact des espèces invasives sur les écosystèmes.

B. Perspectives

Si de nombreux paramètres et phénomènes qui interviennent dans la distribution d'une espèce ne permettent pas de décrire parfaitement les niches (observée, réalisée, fondamentale…) des espèces, l'amélioration des méthodes utilisées et leur sélection en fonction de l'objectif de

l'étude devraient permettre d'en augmenter l'efficacité. J'ai donc regroupé ci-dessous quelques points d'étude qui me semblent prioritaires pour optimiser l'utilisation des modèles de niche.

- **Détermination précise des conditions d'application de la méthode itérative**

Nous avons vu, aussi bien avec des espèces virtuelles qu'avec des espèces réelles que la méthode itérative permettait d'améliorer la qualité des prédictions dans le cas d'absences environnementales abondantes. Il faudrait poursuivre les études pour établir un cadre rigoureux pour l'utilisation de cette méthode. De nombreux points restent à étudier :

- nombre minimal et qualité des observations utilisées. La capacité de la méthode itérative à reconstruire la niche de l'espèce dépend non seulement du nombre d'occurrences disponibles mais aussi de leur répartition dans l'espace environnemental (regroupement dans certaines zones ou au contraire large couverture des différentes conditions possibles) ;
- choix de la valeur seuil. Dans nos études, nous avons utilisé la maximisation du kappa dans le but d'éviter un accroissement « incontrôlé » de la prévalence prédite, les critères basés sur la sensitivité et la spécificité ayant tendance à surestimer la distribution. Mais il est possible que dans le cas d'absences très nombreuses, le choix du kappa comme valeur seuil, en respectant trop la prévalence observée, limite l'efficacité de la méthode itérative. L'utilisation d'autres valeurs seuils pourrait donc s'avérer plus efficace dans certains cas ;
- biais liés aux fausses présences. Les fausses présences sont généralement rares et donc peu étudiées dans le cadre des modèles de niche. Cependant, la nature du modèle itératif pourrait augmenter sa sensibilité à ce type d'erreurs de la base de données de calibration ;

- sensibilité au choix des variables climatiques. Comme pour les fausses présences, la nature itérative de la méthode pourrait aggraver les conséquences du choix de mauvaises variables environnementales.

- **Evaluation de la qualité des modèles**

A l'heure actuelle, la mesure de la qualité des modèles est essentiellement basée sur l'utilisation de quelques indices (AUC, Kappa, TSS) qui présentent tous des biais identifiés (influence de la prévalence, de la spécialisation de l'espèce, de l'étendue de l'aire d'étude, même poids donné aux deux types d'erreurs…). L'utilisation de ces indices est d'autant plus biaisée que certains servent également à la détermination de la valeur seuil. On n'a donc pas d'indépendance entre les prédictions et les mesures de qualité. Il existe également souvent une dépendance entre la base de calibration et la base de test, celles-ci étant issues de la même zone géographique. Et si l'utilisation de deux zones géographiques différentes permet de s'affranchir de la corrélation spatiale, elle peut conduire à une perte d'information. Si de nombreuses études se sont penchées sur ces problématiques (Randin et al. 2006; Mouton et al. 2010), des efforts restent à faire, en particulier pour prendre en compte la localisation géographique des erreurs, la présence de fausses absences dans la base de test et le choix des mesures de qualité en fonction de l'objectif de l'étude.

- **Prise en compte des contraintes de dispersion et des activités anthropiques**

Les mouvements des espèces ne se font pas au hasard. Si la plupart des arrivées d'espèces invasives de poissons ont l'homme pour vecteur, certaines se font par dispersion naturelle, soit à partir de la zone native par des voies d'eaux récemment créées (canaux…) soit, au sein de la zone exotique, par le réseau naturel.

L'utilisation simultanée des résultats des modèles de niche, qui identifient les zones d'établissement potentiel des espèces exotiques et de données d'activité anthropique (flux commerciaux, niveau de vie…) correspondant aux régions les plus à risque d'introduction

(Leprieur et al. 2008), devrait permettre de cibler les zones devant faire l'objet de mesures prioritaires dans le cadre de la lutte contre les espèces invasives. L'intégration des contraintes de dispersion naturelle dans les modèles, déjà largement envisagée pour évaluer la distribution future des espèces natives sous l'effet du changement climatique (Franklin 2010; Buse and Griebeler 2011), peut aussi être utile dans le cadre des espèces invasives. Elle permet d'évaluer la dynamique du processus d'invasion (Smolik et al. 2010) et de prendre en compte l'effet des changements environnementaux rapides, en particulier pour les espèces dispersant lentement (Zurell et al. 2009). La connaissance du processus spatio-temporel d'invasion, au moins à des échelles relativement larges, est particulièrement pertinente pour augmenter l'efficacité des mesures de contrôle des invasives déjà identifiées.

La prise en compte de la dispersion peut également se révéler utile pour l'étude des mesures de conservation d'espèces actuellement non invasives dont l'aire de distribution devrait se modifier sous l'effet des changements globaux. En effet, une aide à la dispersion par migration assistée (par exemple l'introduction dans un bassin hydrographique en bordure de l'aire native) a été envisagée pour certaines espèces. Mais une fois introduite dans ce nouvel environnement, l'espèce, jusque là en danger, peut se révéler invasive (Loss et al. 2011). La prise en compte des contraintes de dispersion, en permettant de déterminer la distribution finale de l'espèce après l'épisode de migration assistée, aiderait à cibler les zones préférentielles d'introduction, permettant la préservation de l'espèce sans générer une nouvelle espèce invasive.

- **Introduction de nouveaux paramètres dans les modèles**

Les modèles que nous avons utilisés pour prédire les risques d'invasion des espèces exotiques de poissons prennent en compte, en plus des variables climatiques, des variables topographiques décrivant la position des sites le long du gradient amont-aval et la vélocité du

courant. Ces paramètres se sont révélés essentiels pour la modélisation de la niche des poissons (Buisson et al. 2008). Si le changement climatique devrait avoir peu d'influence sur la position dans le gradient, sauf dans des cas très particuliers, il risque d'affecter largement l'hydrologie des cours d'eau en modifiant la pluviométrie du bassin versant. La qualité des prédictions sera donc améliorée par l'intégration de prédictions hydrologiques. Malheureusement, les prédictions disponibles actuellement montrent une très grande variabilité liée aux modèles utilisés (Döll and Zhang 2010). Leur incorporation dans les modèles de distribution pourrait conduire à l'introduction d'une nouvelle source d'incertitude dans les prédictions et nécessite donc des études avant une utilisation à large échelle.

L'occupation des sols peut également avoir un impact sur la distribution des espèces invasives de poissons en changeant le régime hydrologique des cours d'eau mais aussi en modifiant la qualité de l'eau, d'autant que certaines espèces invasives sont présentes préférentiellement dans les sites perturbés (Marchetti et al. 2004a). L'incorporation de l'occupation des sols et de ses changements dans le futur, déjà utilisée pour modéliser la distribution d'autres taxa (Franklin 2010; Jiguet et al. 2010), devrait également permettre d'améliorer la prédiction des risques d'invasion des milieux dulçaquicoles.

L'introduction de variables permettant de prendre en compte les perturbations d'origine anthropique qui créent de nouvelles niches et permettent de mesurer la pression de propagule est d'autant plus importante que la distribution des espèces invasives semble dans certains cas plus liée à celles-ci qu'aux variables climatiques (Marini et al. 2012).

L'amélioration des modèles de niche, en augmentant la robustesse et la qualité des prédictions, et la prise en compte des facteurs d'introduction, en identifiant les régions les plus à risques, devraient permettre d'améliorer la mise en place de stratégies de prévention et de lutte contre les invasions biologiques.

Glossaire

Absence : site (ou point d'échantillonnage dans une base de données) où l'espèce n'a pas été observée. L'absence d'une espèce peut avoir de nombreuses origines qui impactent différemment les modèles de niche.

Absence contingente : l'absence de l'espèce est due à des facteurs non environnementaux (barrière à la dispersion, interactions biotiques, extirpations).

Absence environnementale : l'espèce est absente du site car les conditions environnementales ne lui permettent pas de maintenir des populations viables.

Absence méthodologique : l'espèce est présente sur le site mais elle n'a pas été observée lors de l'échantillonnage.

Aire native : zone géographique où l'espèce était naturellement présente avant intervention humaine.

Aire exotique : zone géographique où l'espèce est présente suite à des introductions.

AUC : indice de qualité des modèles de distribution de présence/absence basé sur les probabilités de présence de l'espèce.

Détectabilité : probabilité qu'une espèce qui est présente en un site soit observée lors de l'échantillonnage.

Etablissement : l'espèce, jusque là présente suite à des introductions, maintient des populations viables sur plusieurs générations sans nouvelles phases d'introduction.

Fausses absences : le modèle de niche utilisé prédit l'absence de l'espèce dans des sites où elle est présente (ou des sites où elle pourrait s'établir).

Fausses présences : le modèle de niche utilisé prédit la présence de l'espèce dans des sites où elle est absente (ou des sites où elle ne peut pas s'établir).

Introduction : une espèce est relâchée (volontairement ou involontairement) par l'homme dans un milieu où elle était jusque là absente.

Invasion : une espèce établie accroît largement ses populations et son aire de répartition en impactant les écosystèmes receveurs.

Kappa : indice de mesure de la qualité des modèles de présence/absence basé sur les prédictions de présence et absence et prenant en compte la possibilité de prédire correctement les sites sous l'effet du hasard.

Modèles corrélatifs : modèles déterminant la distribution de l'espèce en établissant un lien statistique entre les données de présence ou de présence/absence de l'espèce et les variables environnementales.

Modèle de niche « présence only » : modèle corrélatif basé uniquement sur les données environnementales correspondant aux sites où l'espèce a été observée (et sur le « fond » des conditions observées dans l'ensemble de la zone d'étude). Dans ce type de modèle, les sites échantillonnés où l'espèce n'a pas été détectée ne sont pas pris en compte.

Modèle de niche présence/absence : modèle corrélatif basé sur les données correspondant à la fois aux sites où l'espèce a été observée et à des sites d'absence. Ces derniers peuvent être des sites échantillonnés où l'espèce n'a pas été observée ou des sites pris au hasard dans la zone d'étude et où on suppose l'espèce absente (pseudo-absences).

Modèles hybrides : modèles déterminant la distribution de l'espèce en couplant une approche mécanistique et une approche corrélative.

Modèles mécanistiques : modèles basés sur les caractéristiques physiologiques et écologiques de l'espèce pour déterminer la distribution d'une espèce.

Niche fondamentale : région de l'espace des variables environnementales correspondant aux conditions qui permettraient à l'espèce de maintenir des populations viables.

Niche réalisée : région de l'espace des variables environnementales correspondant aux conditions dans lesquelles on trouve l'espèce.

Occurrence : site ou point de la base de données où l'espèce a été observée.

Présence : site ou point de la base de données où l'espèce a été observée. Si ces données sont généralement fiables d'un point de vue écologique, la présence d'une espèce peut parfois être accidentelle (mouvement entre deux zones géographiques de la niche…), l'espèce peut également avoir été introduite mais ne pas s'être établie ou l'observation peut être due à une erreur d'identification de l'espèce.

Pression de propagule : mesure du nombre d'individus introduits et du nombre de tentatives d'introductions de l'espèce exotique dans un lieu.

Prévalence : proportion des sites où l'espèce a été observée.

Sensitivité : proportion des sites de présence de l'espèce correctement prédits par le modèle de niche (présence/absence).

Spécificité : proportion des sites d'absence de l'espèce correctement prédits par le modèle de niche (présence/absence).

TSS : indice de mesure de la qualité des modèles de présence/absence basé sur les prédictions de présence et absence et égal à la somme de la sensitivité et de la spécificité moins 1.

Phénologie : calendrier des évènements périodiques chez les êtres vivants liés aux variations climatiques saisonnières (floraison, migration…).

Pseudo-absences : données utilisées dans les modèles corrélatifs de présence/absence et correspondant à des sites de la zone d'étude non échantillonnés où l'on suppose l'espèce absente.

Références

A

Allan, J. D., R. Abell, Z. Hogan, C. Revenga, B. W. Taylor, R. L. Welcomme, and K. Winemiller. 2005. Overfishing of inland waters. Bioscience 55:1041-1051.

Allouche, O., A. Tsoar, and R. Kadmo. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43:1223-1232.

Anderson, R. P., and I. Gonzalez Jr. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions: An implementation with Maxent. Ecological Modelling 222 2796– 2811.

Araújo, M. B., and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47.

Araújo, M. B., W. Thuiller, and R. G. Pearson. 2006. Climate warming and the decline of amphibians and reptiles in Europe. Journal of Biogeography 33:1712-1728.

Araújo, M. B., R. J. Whittaker, R. J. Ladle, and M. Erhard. 2005. Reducing uncertainty in projections of extinction risk from climate change. Global Ecology and Biogeography 14:529–538.

Austin, M. P., and K. P. Van Niel. 2011. Improving species distribution models for climate change studies: variable selection and scale. Journal of Biogeography 38:1–8.

B

Barbet-Massin, M., B. A. Walther, W. Thuiller, C. Rahbek, and F. Jiguet. 2009. Potential impacts of climate change on the winter distribution of Afro-Palaearctic migrant passerines. Biology Letters 5:248-251.

Baxter, C. V., K. D. Fausch, M. Murakami, and P. L. Chapman. 2004. Fish invasion restructures stream and forest food webs by interrupting reciprocal prey subsidies. Ecology 85:2656-2663.

Beaumont, L. J., R. V. Gallagher, W. Thuiller, P. O. Downey, M. R. Leishman, and L. Hughes. 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Diversity and Distributions 15:409-420.

Bellard, C., C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp. 2012. Impacts of climate change on the future of biodiversity. Ecology Letters doi: 10.1111/j.1461-0248.2011.01736.x.

Bianco, P. G., and P. Turin. 2010. Record of two established populations of nile tipalia, *Oreochromis niloticus*, in freshwater of northern Italy. Journal of Applied Ichthyology 26:140-142.

Blackburn, T. M., and R. P. Duncan. 2001. Determinants of establishment success in introduced birds. Nature 414:195-197.

Blanchet, S., G. Loot, G. Grenouillet, and S. Brosse. 2007. Competitive interactions between native and exotic salmonids: a combined field and laboratory demonstration. Ecology of Freshwater Fish 16:133-143.

Bombi, P., L. Luiselli, and M. D'Amen. 2011. When the method for mapping species matters: defining priority areas for conservation of African freshwater turtles. Diversity and Distribution 17:581–592.

Bomford, M., F. Kraus, S. C. Barry, and E. Lawrence. 2009. Predicting establishment success for alien reptiles and amphibians: a role for climate matching. Biological Invasions 11:713-724.

Bond, N., J. Thomson, P. Reich, and J. Stein. 2011. Using species distribution models to infer potential climate change-induced range shifts of freshwater fish in south-eastern Australia. Marine and Freshwater Research 62:1043-1061.

Braunisch, V., and R. Suchant. 2010. Predicting species distributions based on incomplete survey data: the trade-off between precision and scale. Ecography 33:826-840.

Breiman, L. 2001. Random Forests. Machine Learning 45:5-32.

Broennimann, O., and A. Guisan. 2008. Predicting current and future biological invasions: both native and invaded ranges matter. Biology Letters 4:585-589.

Brown, J. H. 1984. On the relationship between abundance and distribution species. The American Naturalist 124:225-279.

Buisson, L., L. Blanc, and G. Grenouillet. 2008. Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors. Ecology of Freshwater Fish 17:244-257.

Buisson, L., W. Thuiller, N. Casajus, S. Lek, and G. Grenouillet. 2010. Uncertainty in ensemble forecasting of species distribution. Global Change Biology 16:1145-1157.

Buse, J., and E. M. Griebeler. 2011. Incorporating classified dispersal assumptions in predictive distribution models – A case study with grasshoppers and bush-crickets 222 2130–2141.

C

Canonico, G. C., A. Arthington, J. K. McCrary, and M. L. Thieme. 2005. The effects of introduced tilapias on native biodiversity. Aquatic Conservation-Marine and Freshwater Ecosystems 15:463-483.

Capinha, C., B. Leung, and P. Anastácio. 2011. Predicting worldwide invasiveness for four major problematic decapods: an evaluation of using different calibration sets. Ecography 34:448-459.

Casal, C. M. V. 2006. Global documentation of fish introductions: the growing crisis and recommendations for action. Biological Invasions 8:3-11.

Cassey, P., T. M. Blackburn, S. Sol, R. P. Duncan, and J. L. Lockwood. 2004. Global patterns of introduction effort and establishment success in birds. Proceedings of The Royal Society of London Series B-Biological Sciences 271:S405-S408.

Chu, C., N. E. Mandrak, and C. K. Minns. 2005. Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. Diversity and Distributions 11:299-310.

Cianfrani, C., G. Le Lay, A. H. Hirzel, and A. Loy. 2010. Do habitat suitability models reliably predict the recovery areas of threatened species? Journal of Applied Ecology 47:421-430.

Clavero, M., and E. Garcia-Berthou. 2005. Invasive species are a leading cause of animal extinctions. Trends In Ecology and Evolution 20:110-110.

Clavero, M., and E. Garcia-Berthou. 2006. Homogenization dynamics and introduction routes of invasive freshwater fish in the Iberian Peninsula. Ecological Applications 16:2313-2324.

Closs, G. P., and P. S. Lake. 1996. Drought, differential mortality and the coexistence of a native and an introduced fish species in a south east Australian intermittent stream. Environmental Biology Of Fishes 47:17-26.

Conrad, J. L., K. L. Weinersmith, T. Brodin, J. B. Saltz, and A. Sih. 2011. Behavioural syndromes in fishes: a review with implications for ecology and fisheries management. Journal Of Fish Biology 78:395-435.

Crawford, S. S., and A. M. Muir. 2008. Global introductions of salmon and trout in the genus Oncorhynchus: 1870-2007. Reviews in Fish Biology and Fisheries 18:313-344.

Crossman, N. D., B. A. Bryan, and D. M. Summers. 2012. Identifying priority areas for reducing species vulnerability to climate change. Diversity and Distribution 18:60-72.

Cubaynes, S., R. Pradel, R. Choquet, C. Duchamp, J.-M. Gaillard, J.-D. Lebreton, E. Marboutin et al. 2010. Importance of accounting for detection heterogeneity when estimating abundance: the case of french wolves. Conservation Biology 24:621-626.

Cucherousset, J., and J. D. Olden. 2011. Ecological Impacts of Non-native Freshwater Fishes. Fisheries 36:215-230.

Cudmore, B., and N. E. Mandrak. 2004. Biological Synopsis of Grass Carp (*Ctenopharyngodon idella*), Pages 52, Canadian Manuscript Report of Fisheries and Aquatic Sciences 2705.

D

Deacon, A. E., I. W. Ramnarine, and A. E. Magurran. 2011. How reproductive ecology contributes to the spread of a globally invasive fish. PLoS One doi:10.1371/journal.pone.0024416.

DeVaney, S.C. , M. Kristina, J.B. McNyset, W.A.T.Peterson and E.O. Wiley. 2009. A tale of Four ‘‘Carp": invasion potential and ecological niche modeling. Plos One 4. e5451. doi:10.1371/journal.pone.0005451

Dobrowski, S. Z. 2011. A climatic basis for microrefugia: the influence of terrain on climate. Global Change Biology 17:1022-1035.

Döll, P., and J. Zhang. 2010. Impact of climate change on freshwater ecosystems: a global-scale analysis of ecologically relevant river flow alterations. Hydrology and Earth System Sciences 14:783–799.

Domisch, S., S. Jähnig, and P. Haase. 2011. Climate-change winners and losers: stream macroinvertebrates of a submontane region in Central Europe Freshwater Biology 56:2009–2020.

E-F

Elith, J., M. Kearney, and S. Phillips. 2010. The art of modelling range-shifting species. Methods in Ecology and Evolution.

Engler, R., A. Guisan, and L. Rechsteiner. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. Journal of Applied Ecology 41:263-274.

Evans, M. R., K. J. Norris, and T. G. Benton. 2012. Predictive ecology: systems approaches. Philosophical Transactions of the Royal Society B-Biological Sciences 367:163-169.

Fausch, K. D., Y. Taniguchi, S. Nakano, G. D. Grossman, and C. R. Townsend. 2001. Flood disturbance regimes influence rainbow trout invasion success among five Holarctic regions. Ecological Applications 11:1438-1455.

Feeley, K. J., and Silman, M.R. 2011. Keep collecting: accurate species distribution modelling requires more collections than previously thought. Diversity and Distribution 17:1132–1140.

Fitzpatrick, M. C., J. F. Weltzin, N. J. Sanders, and R. R. Dunn. 2007. The biogeography of prediction error: why does the introduced range of the fire ant over-predict its native range? Global Ecology and Biogeography 16:24-33.

Franklin, J. 2009, Mapping species distributions: spatial inference and prediction. Cambridge, Cambridge University Press.

Franklin, J. 2010. Moving beyond static species distribution models in support of conservation biogeography. Diversity and Distributions 16:321-330.

Freeman, E. A., and G. G. Moisen. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecological Modelling 217:48-58.

Friedman, J. H. 2001. Greedy function approximation: A gradient boosting machine. Annals of Statistics 29:1189-1232.

G

Gallien, L., R. Douzet, S. Pratte, N. E. Zimmermann, and W. Thuiller. 2012. Invasive species distribution models - How violating the equilibrium assumption can create new insights. Global Ecology and Biogeography.

Gallien, L., T. Munkemuller, C. H. Albert, I. Boulangeat, and W. Thuiller. 2010. Predicting potential distributions of invasive species: where to go from here? Diversity and Distributions 16:331-342.

Garcia-Berthou, E., C. Alcaraz, Q. Pou-Rovira, L. Zamora, G. Coenders, and C. Feo. 2005. Introduction pathways and establishment rates of invasive aquatic species in Europe. Canadian Journal of Fisheries and Aquatic Sciences 62:453-463.

Godsoe, W. 2010. I can't define the niche but I know it when I see it: a formal link between statistical theory and the ecological niche. Oikos 119:53-60.

Grenouillet, G., L. Buisson, N. Casajus, and S. Lek. 2011. Ensemble modelling of species distribution: the effects of geographical and environmental ranges. Ecography 34 9-17.

Guisan, A., C. H. Graham, J. Elith, F. Huettmann, and t. N. S. D. M. Group. 2007. Sensitivity of predictive species distribution models to change in grain size. Diversity and Distributions 13:332-340.

Guo, Q. F., H. Qian, R. E. Ricklefs, and W. M. Xi. 2006. Distributions of exotic plants in eastern Asia and North America. Ecology Letters 9:827-834.

H

Hanley, J. A., and B. J. McNeil. 1982. The Meaning And Use Of The Area Under A Receiver Operating Characteristic (Roc) Curve. Radiology 143:29-36.

Hastie, T. 2006. GAM: Generalized Additive Models. R package version 0.98.

Hastie, T., and R. Tibshirani. 1990, Generalized additive models: Monographs on Statistics and Applied Probabilities. London, UK, Chapman and Hall.

Heino, J., R. Virkkala, and H. Toivonen. 2009. Climate change and freshwater biodiversity: detected patterns, future trends and adaptations in northern regions. Biological Reviews 84:39-54.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965-1978.

Hill, M. P., A. A. Hoffmann, S. Macfadyen, P. A. Umina, 4, and J. Elith. 2012. Understanding niche shifts: using current and historical data to model the invasive redlegged earth mite, Halotydeus destructor. Diversity and Distribution 18:191-203.

Hooper, D. U., F. S. Chapin, J. J. Ewel, A. Hector, S. L. Inchausti, S., J. H. Lawton, D. M. Lodge et al. 2005. Effects of biodiversity on ecosystem functioning: a consensus of current knowledge. Ecological monographs 75:3-23.

Hortal, J., A. Jimenez-Valverde, J. F. Gomez, J. M. Lobo, and A. Baselga. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. Oikos 117:847-858.

Hu, J., and Z. Jiang. 2010. Predicting the potential distribution of the endangered Przewalski's gazelle. Journal of Zoology 282:54-63.

Huber, R., S. Greco, and J. Thorne. 2010. Spatial scale effects on conservation network design: trade-offs and omissions in regional versus local scale planning. Landscape Ecology 25:683–695.

Hutchinson, G. E. 1957. Concluding remarks. Cold Spring Harbor Symposia on Quantitative Biology 22:415-427.

J

Januchowski-Hartley, S., J. VanDerWal, and D. Sydes. 2011. Effective Control of Aquatic Invasive Species in Tropical Australia. Environmental Management 48:568–576.

Jeschke, J. M., and D. L. Strayer. 2005. Invasion success of vertebrates in Europe and North America. Proceedings of the National Academy of Sciences of the United States Of America 102:7198-7202.

Jeschke, J. M., and D. L. Strayer. 2006. Determinants of vertebrate invasion success in Europe and North America. Global Change Biology 12:1608-1619.

Jiguet, F., R. D. Gregory, V. Devictor, R. E. Green, P. Vorisek, A. Van Strien, and D. Couvet. 2010. Population trends of European common birds are predicted by characteristics of their climatic niche. Global Change Biology 16:497-505.

Jiménez-Valverde, A., and J. M. Lobo. 2006. The ghost of unbalanced species distribution data in geographical model predictions. Diversity and Distributions 12:521-524.

Jiménez-Valverde, A., and J. M. Lobo. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence Acta Oecologica 31:361-369.

Jiménez-Valverde, A., J. M. Lobo, and J. Hortal. 2008. Not as good as they seem: the importance of concepts in species distribution modelling. Diversity and Distributions 14:885-890.

Jiménez-Valverde, A., A. T. Peterson, J. J. Soberón, J. M. Overton, P. Aragón, and J. M. Lobo. 2011. Use of niche models in invasive species risk assessments. Biological Invasions 13:2785-2797.

Johnson, D. M., A. M. Liebhold, P. C. Tobin, and O. N. Bjornstad. 2006. Allee effects and pulsed invasion by the gypsy moth. Nature 444:361-363.

K

Kearney, M., B. L. Phillips, C. R. Tracy, K. A. Christian, G. Betts, and W. P. Porter. 2008. Modelling species distributions without using species distributions: the cane toad in Australia under current and future climates. Ecography 31:423-434.

Kearney, M., and W. Porter. 2009. Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecology Letters 12:334-350.

Kearney, M. R., B. A. Wintle, and W. P. Porter. 2010. Correlative and mechanistic models of species distribution provide congruent forecasts under climate change. Conservation Letters 3:203-213.

Kolar, C. S., and D. M. Lodge. 2001. Progress in invasion biology: predicting invaders. Trends in Ecology and Evolution 16:199-204.

Kovacs, K., T. Václavík, R. Haight, A. Pang, N. J. Cunniffe, C. Gilligan, R. K. et al. 2011. Predicting the economic costs and property value losses attributed to sudden oak death damage in California (2010-2020). Journal of Environmental Management 92 1292-1302.

Kriticos, D. J., and A. Leriche. 2010. The effects of climate data precision on fitting and projecting species niche models. Ecography 33:115-127.

L

Larson, E. R., J. D. Olden, and N. Uso. 2010. Decoupled conservatism of Grinnellian and Eltonian niches in an invasive arthropod. Ecosphere 1.

Leprieur, F., O. Beauchard, S. Blanchet, T. Oberdorff, and S. Brosse. 2008. Fish invasions in the world's river systems: when natural processes are blurred by human activities PloS Biology 6:e28. http://dx.doi.org/10.1371/journal.pbio.0060028.

Leprieur, F., S. Brosse, E. Garcia-Berthou, T. Oberdorff, J. D. Olden, and C. R. Townsend. 2009a. Scientific uncertainty and the assessment of risks posed by non-native freshwater fishes. Fish and Fisheries 10:88-97.

Leprieur, F., J. D. Olden, S. Lek, and S. Brosse. 2009b. Contrasting patterns and mechanisms of spatial turnover for native and exotic freshwater fish in Europe. Journal of Biogeography 36:1899-1912.

Liaw, A., and M. Wiener. 2002. Classification and regression by RandomForest. R News 2:18-22.

Liu, C. R., P. M. Berry, T. P. Dawson, and R. G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385-393.

Lobo, J. M., A. Baselga, J. Hortal, A. Jimenez-Valverde, and J. F. Gomez. 2007. How does the knowledge about the spatial distribution of Iberian dung beetle species accumulate over time? Diversity and Distributions 13:772-780.

Lobo, J. M., A. Jiménez-Valverde, and J. Hortal. 2010. The uncertain nature of absences and their importance in species distribution modelling. Ecography 33:103-114.

Lobo, J. M., A. Jimenez-Valverde, and R. Real. 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecology and Biogeography 17:145-151.

Lockwood, J. L., M. F. Hoopes, and M. P. Marchetti. 2007, Invasion Ecology. United Kingdom, Blackwell Scientific Press.

Lomba, A., L. Pellissier, C. Randin, J. Vicente, F. Moreira, J. Honrad, and A. Guisan. 2010. Overcoming the rare species modelling paradox: A novel hierarchical framework applied to an Iberian endemic plant. Biological Conservation 143.

Loss, S. R., L. A. Terwilliger, and A. C. Peterson. 2011. Assisted colonization: Integrating conservation strategies in the face of climate change. Biological Conservation 144:92-100.

M

Mack, R. N., D. Simberloff, W. M. Lonsdale, H. Evans, M. Clout, and F. A. Bazzaz. 2000. Biotic invasions: Causes, epidemiology, global consequences, and control. Ecological Applications 10:689-710.

Manel, S., H. C. Williams, and S. J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. Journal of Applied Ecology 38:921-931.

Marchetti, M. P., T. Light, P. B. Moyle, and J. H. Viers. 2004a. Fish invasions in California watersheds: Testing hypotheses using landscape patterns. Ecological Applications 14:1507-1525.

Marchetti, M. P., P. B. Moyle, and R. Levine. 2004b. Invasive species profiling? Exploring the characteristics of non-native fishes across invasion stages in California. Freshwater Biology 49:646-661.

Marini, L., A. Battisti, E. Bona, G. Federici, F. Martini, M. Pautasso, and P. E. Hulme. 2012. Alien and native plant life-forms respond differently to human and climate pressures Global Ecology and Biogeography 21:534–544.

Marmion, M., M. Parviainen, M. Luoto, R. K. Heikkinen, and W. Thuiller. 2009. Evaluation of consensus methods in predictive species distribution modelling. Diversity and Distributions 15:59-69.

Maron, J. L., M. VilÃ , R. Bommarco, S. Elmendorf, and P. Beardsley. 2004. Rapid Evolution of an Invasive Plant. Ecological Monographs 74:261-280.

Mateo-Tomas, P., and P. P. Olea. 2009. Combining scales in habitat models to improve conservation planning in an endangered vulture. Acta Oecologica-International Journal of Ecology 35:489-498.

McCullagh, P., and J. Nelder. 1989, Generalized linear models. New York,USA, Chapman and Hall.

McInnes, L., A. Purvis, and C. D. L. Orme. 2009. Where do species' geographic ranges stop and why? Landscape impermeability and the Afrotropical avifauna. Proceedings of the Royal Society B, Biological Sciences 276:3063-3070.

Medley, K. A. 2010. Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. Global Ecology and Biogeography 19:122-133.

Meffe, G. K. 1991. Life-history changes in eastern mosquitofish (*Gambusia holbrooki*) induced by thermal elevation. Canadian Journal of Fisheries and Aquatic Sciences 48:60-66.

Meffe, G. K., S. C. Weeks, M. Mulvey, and K. L. Kandl. 1995. Genetic differences in thermal tolerance of eastern mosquitofish (*Gambusia holbrooki* Poeciliidae) from ambient and thermal ponds. Canadian Journal of Fisheries and Aquatic Sciences 52:2704-2711.

Menke, S. B., D. A. Holway, R. N. Fisher, and W. Jetz. 2009. Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. Global Ecology and Biogeography 18:50-63.

Millar, C. S., and G. Blouin-Demers. 2012. Habitat suitability modelling for species at risk is sensitive to algorithm and scale: A case study of Blanding's turtle, *Emydoidea blandingii*, in Ontario, Canada. Journal for Nature Conservation 20:18-29.

Millennium Ecosystem Assessment. 2005. Ecosystems and human well being: Biodiversity Synthesis, World Resources Institute, Washington, DC.

Minns, C. K., and J. E. Moore. 1995. Factors limiting the distributions of Ontario's freshwater fish: the role of climate and other variables, and the potential impacts of climate change. Canadian Journal of Fisheries and Aquatic Sciences 121:137-160.

Mitchell, A. L., and J. H. Knouft. 2009. Non-native fishes and native species diversity in freshwater fish assemblages across the United States. Biological Invasions 11:1441-1450.

Monti, D., P. Keith, and E. Vigneux. 2010, Atlas des poissons et des crustacé d'eau douce de: Patrimoines Naturels. Paris, SPN / IEGB / MNHN.

Morin, X., and W. Thuiller. 2009. Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. Ecology 90:1301-1313.

Moritz, C., J. L. Patton, C. J. Conroy, J. L. Parra, G. C. White, and S. R. Beissinger. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322:261-264.

Mouton, A. M., B. De Baets, and P. L. M. Goethalsa. 2010. Ecological relevance of performance criteria for species distribution models. Ecological Modelling 221:1995-2002.

Moyle, P. B., and T. Light. 1996. Biological invasions of fresh water: empirical rules and assembly theory. Biological Conservation 78:149-161.

Moyle, P. B., and M. P. Marchetti. 2006. Predicting invasion success: Freshwater fishes in California as a model. Bioscience 56:515-524.

Müller, U. C., J. Pross, P. C. Tzedakis, C. Gamble, U. Kotthoff, G. Schmiedl, S. Wulf et al. 2011. The role of climate in the spread of modern humans into Europe. Quaternary Science Reviews 30:273-279.

N

Nenzen, H. K., and M. B. Araujo. 2011. Choice of threshold alters projections of species range shifts under climate change. Ecological Modelling 222:3346-3354.

Newbold, T., T. Reader, A. El-Gabbas, W. Berg, W. M. Shohdi, S. Zalat, S. B. El Din et al. 2010. Testing the accuracy of species distribution models using species records from a new field survey. Oikos 119:1326-1334.

Nilsson, C., C. A. Reidy, M. Dynesius, and C. Revenga. 2005. Fragmentation and flow regulation of the world's large river systems. Science 308:405-408.

P

Parmesan, C., N. Ryrholm, C. Stefanescu, J. K. Hill, C. D. Thomas, H. Descimon, B. Huntley et al. 1999. Poleward shifts in geographical ranges of butterfly species associated with regional warming. Nature 399:579-583.

Parmesan, C., and G. Yohe. 2003. A globally coherent fingerprint of climate change impacts across natural systems. Nature 421:37-42.

Pearson, R. G., C. J. Raxworthy, M. Nakamura, and A. T. Peterson. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. Journal of Biogeography 34:102-117.

Peterson, A. T. 2011. Ecological niche conservatism: a time-structured review of evidence. Journal of Biogeography 38:817–827.

Peterson, A. T., M. Papes, and J. Soberon. 2008. Rethinking receiver operating characteristic analysis applications in ecological niche modeling. Ecological Modelling 213:63-72.

Petitpierre, B., C. Kueffer, O. Broennimann, C. Randin, C. Daehler, and A. Guisan. 2012. Climatic niche shifts are rare among terrestrial plant invaders. Science 335:1344-1348.

Phillips, S. J., M. Dudik, J. Elith, C. H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. 2009. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19:181-197.

Pimentel, D., L. Lach, R. Zuniga, and D. Morrison. 2000. Environmental and economic costs of nonindigenous species in the United States. Bioscience 50:53-65.

Pimentel, D., R. Zuniga, and D. Morrison. 2005. Update on the Environmental and Economic Costs Associated with Alien-Invasive Species in the United States. Ecological Economics 52: 273-288.

Pysek, P., V. Jarosik, P. E. Hulme, I. Kuhn, J. Wild, M. Arianoutsou, S. Bacher et al. 2010. Disentangling the role of environmental and human pressures on biological invasions across Europe. Proceedings of the National Academy of Sciences 107:12157-12162.

Pysek, P., D. M. Richardson, J. Pergl, V. Jarosík, Z. Sixtová, and E. Weber. 2008. Geographical and taxonomic biases in invasion ecology. Trends in Ecology and Evolution 23:237- 244.

R

Rahel, F. J. 2007. Biogeographic barriers, connectivity and homogenization of freshwater faunas: it's a small world after all. Freshwater Biology 52:696-710.

Rahel, F. J., and J. D. Olden. 2008. Assessing the Effects of Climate Change on Aquatic Invasive Species. Conservation Biology 22:521–533.

Randin, C. F., T. Dirnbo, S. Dullinger, N. E. Zimmermann, M. Zappa, and A. Guisan. 2006. Are niche-based species distribution models transferable in space? Journal of Biogeography 33:1689–1703.

Richardson, D. M., P. Pyšek, M. Rejmánek, M. G. Barbour, F. D. Panetta, and C. J. West. 2000. Naturalization and invasion of alien plants: concepts and definitions. Diversity and Distributions 6:93–107.

Richter, B. D., D. P. Braun, M. A. Mendelson, and L. L. Master. 1997. Threats to imperiled freshwater fauna. Conservation Biology 11:1081-1093.

Root, T. L., J. T. Price, K. R. Hall, S. H. Schneider, and J. A. Pounds. 2003. Fingerprints of global warming on wild animals and plants. Nature 421:57-60.

Rowe, R. J., J. A. Finarelli, and E. A. Rickart. 2010. Range dynamics of small mammals along an elevational gradient over an 80-year interval. Global Change Biology 16:2930-2943.

Ruesink, J. L. 2005. Global analysis of factors affecting the outcome of freshwater fish introductions. Conservation Biology 19:1883-1893.

S

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Global Ecology and Biogeography 20:181-192.

Sauer, J., S. Domisch, C. Nowak, and P. Haase. 2011. Low mountain ranges: summit traps for montane freshwater species under climate change Biodiversity and Conservation 13:3133- 3146.

Schlaepfer, M. A., P. W. Sherman, B. Blossey, and M. C. Runge. 2005. Introduced species as evolutionary traps. Ecology Letters 8:241-246.

Schulte, U., A. Hochkirch, S. Lötters, D. Rödder, S. Schweiger, T. Weimann, and M. Veith. 2011. Cryptic niche conservatism among evolutionary lineages of an invasive lizardg. Global Ecology and Biogeography 665:198-211.

Scordella, G., F. Lumare, A. Conides, and C. Papaconstantinou. 2003. First occurrence of the tilapia *Oreochromis niloticus niloticus* (Linnaeus, 1758) in Lesina Lagoon (eastern Italian

coast). Mediterranean Marine Science 4:41-47.

Segurado, P., J. M. Santosa, D. Pontb, A. H. Melcherc, D. G. Jalond, R. M. Hughese, and M. T. Ferreira. 2011. Estimating species tolerance to human perturbation: Expert judgment versus empirical approaches. Ecological Indicators 11 1623–1635.

Seo, C., J. H. Thorne, L. Hannah, and W. Thuiller. 2009. Scale effects in species distribution models: implications for conservation planning under climate change. Biology Letters 5:39-43.

Sexton, J. P., J. K. McKay, and A. Sala. 2002. Plasticity and genetic diversity may allow saltcedar to invade cold climates in North America. Ecological Applications 12:1652-1660.

Sharma, S., and D. A. Jackson. 2008. Predicting smallmouth bass (Micropterus dolomieu) occurrence across North America under climate change: a comparison of statistical approaches. Canadian Journal of Fisheries and Aquatic Sciences 65:471-481.

Shelton, J. M., J. A. Day, and C. L. Griffiths. 2008. Influence of largemouth bass, Micropterus salmoides, on abundance and habitat selection of Cape galaxias, *Galaxias zebratus*, in a mountain stream in the Cape Floristic Region, South Africa. African Journal of Aquatic Science 33:201-210.

Shireman, J. V., and C. R. Smith. 1983. Synopsis of biological data on the grass carp, *Ctenopharyngodon idella* (Cuvier et Valenciennes, 1844). Pages 86, Food and Aquaculture Organization Synopsis.

Simberloff, D., and P. Stiling. 1996. How risky is biological control? Ecology 77:1965-1974.

Smolik, M. G., S. Dullinger, F. Essl, I. Kleinbauer, M. Leitner, J. Petrerseil, L.-M. Stadler et al. 2010. Integrating species distribution models and interacting particle systems to predict the spread of an invasive alien plant. Journal of Biogeography 37:411-422.

Soberón, J., and A. T. Peterson. 2005. Interpretation of models of fundamental ecological niches and species' distributional areas. Biodiversity Informatics 2:2-10.

Stankowski, P. A., and W. H. Parker. 2011. Future distribution modelling: A stitch in time is not enough. Ecological Modelling 222:567–572.

Stohlgren, T. J., P. Ma, S. Kumar, M. Rocca, J. T. Morisette, C. S. Jarnevich, and N. Benson. 2010. Ensemble Habitat Mapping of Invasive Plant Species. Risk Analysis 30:224-235.

Synes, N. W., and P. E. Osborne. 2011. Choice of predictor variables as a source of uncertainty in continental-scale species distribution modelling under climate change. Global Ecology and Biogeography 20:904–914.

Syphard, A. D., and J. Franklin. 2009. Differences in spatial predictions among species distribution modeling methods vary with species traits and environmental predictors. Ecography 32:907-918.

T

Taylor, B. W., and R. E. Irwin. 2004. Linking economic activities to the distribution of exotic plants. Proceedings of the National Academy of Sciences of the United States of America 101:17725-17730.

Therneau, T. M., and B. Atkinson. 2007. rpart : Recursive Partitioning. R package version 3.1-38.

Thuiller, W., L. Brotons, M. B. Araújo, and S. Lavorel. 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography 27:165-172.

Thuiller, W., S. Lavorel, M. B. Araújo, M. T. Sykes, and I. C. Prentice. 2005a. Climate change threats to plant diversity in Europe. Proceedings of the National Academy of Sciences of the United States of America 102:8245-8250.

Thuiller, W., D. M. Richardson, P. Pyšek, G. F. Midgley, G. O. Hughes, and M. Rouget. 2005b. Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. Global Change Biology 11:2234-2250.

Townsend, C. R. 1996. Invasion biology and ecological impacts of brown trout *Salmo trutta* in New Zealand. Biological Conservation 78:13-22.

Townsend, C. R., and T. A. Crowl. 1991. Fragmented population structure in a native New Zealand fish: an effect of introduced brown trout? Oikos 61:348–354.

U-V

Urban, M. C., B. L. Phillips, D. K. Skelly, and R. Shine. 2007. The cane toad's (*Chaunus [Bufo] marinus*) increasing ability to invade Australia is revealed by a dynamically updated range model. Proceedings of the Royal Society B-Biological Sciences 274:1413-1419.

Václavík, T., and R. K. Meentemeyer. 2011. Equilibrium or not? Modelling potential distribution of invasive species in different stages of invasion. Diversity and Distribution 18:73- 83.

Vander Zanden, M. J. 2005. The success of animal invaders. Proceedings of The National Academy of Sciences of the United States of America 102:7055-7056.

Venables, W. N., and B. D. Ripley. 2002, Modern Applied Statistics with S, Springer-Verlag.

Villemant, C., M. Barbet-Massin, A. Perrard, F. Muller, O. Gargominy, F. Jiguet, and Q. Rome. 2011. Predicting the invasion risk by the alien bee-hawking Yellow-legged hornet *Vespa velutina nigrithorax* across Europe and other continents with niche models. Biological Conservation 144:2142–2150.

Vitousek, P. M., C. M. Dantonio, L. L. Loope, and R. Westbrooks. 1996. Biological invasions as global environmental change. American Scientist 84:468-478.

Vitousek, P. M., H. A. Mooney, J. Lubchenco, and J. M. Melillo. 1997. Human domination of Earth's ecosystems. Science 277:494-499.

W

Warren, D. L., and S. N. Seifert. 2011. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. Ecological Applications 21:335–342.

Webber, B. L., C. J. Yates, D. C. Le Maitre, J. K. Scott, D. J. Kriticos, N. Ota, A. McNeill et al. 2011. Modelling horses for novel climate courses: insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. Diversity and Distribution 17:978–1000.

Weyl, P. S. R., F. C. de Moor, M. P. Hill, and O. L. F. Weyl. 2010. The effect of largemouth bass Micropterus salmoides on aquatic macro-invertebrate communities in the Wit River, Eastern Cape, South Africa. African Journal of Aquatic Science 35:273-281.

Wiens, J. A., and D. Bachelet. 2009. Matching the multiple scales of conservation with the multiple scales of climate change. Conservation Biology 24:51-62.

Wilcove, D. S., D. Rothstein, J. Dubow, A. Phillips, and E. Losos. 1998. Quantifying threats to imperiled species in the United States. Bioscience 48:607-615.

Williams, J. N., C. Seo, J. Thorne, J. K. Nelson, S. Erwin, J. M. O'Brien, and M. W. Schwartz. 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions 15:565-576.

Williamson, M. 1996, Biological Invasions. London, Chapman & Hall.

Wilson, J. R. U., D. M. Richardson, M. Rouget, S. Proches, M. A. Amis, L. Henderson, and W. Thuiller. 2007. Residence time and potential range: crucial considerations in modelling plant invasions. Diversity and Distributions 13:11-22.

Wisz, M. S., and A. Guisan. 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecology 9:8.

Wonham, M. J., J. T. Carlton, G. M. Ruiz, and L. D. Smith. 2000. Fish and ships: relating dispersal frequency to success in biological invasions. Marine Biology 136:1111-1121.

X-Y-Z

Xenopoulos, M. A., and D. M. Lodge. 2006. Going with the flow: Using species-discharge relationships to forecast losses in fish biodiversity. Ecology 87:1907-1914.

Yang, C.-C., M. S. Ascunce, L.-Z. Luo, J. G. Shao, C.-J. Shih, and D. Shoemaker. 2012. Propagule pressure and colony social organization are associated with the successful invasion and rapid range expansion of fire ants in China. Molecular ecology 21:817-833.

Zambrano, L., E. Martinez-Meyer, N. Menezes, and A. T. Peterson. 2006. Invasive potential of common carp (*Cyprinus carpio*) and Nile tilapia (*Oreochromis niloticus*) in American freshwater systems. Canadian Journal of Fisheries and Aquatic Sciences 63:1903-1910.

Zurell, D., F. Jeltsch, C. F. Dormann, and B. Schroder. 2009. Static species distribution models in dynamically changing systems: how good can predictions really be? Ecography 32:733-744.

Liste des figures

Manuscrits

Liste des manuscrits

(M1) Geometry drives the grain size effects in species distribution

En revision dans Ecography

(M2) Identifying climatic niche shifts using coarse-grained occurrence data: a test with non-native freshwater fish.

Global Ecology and Biogeography 20: 407-414

(M3) Dealing with noisy absences to optimize species distribution models: an iterative ensemble modelling approach.

Soumis à PLOS One

(M4) The iterative ensemble modelling approach increases the accuracy of fish distribution models

En préparation

Geometry drives the grain size effects in species distribution models

Christine LAUZERAL[1, 2]**, Gaël GRENOUILLET**[1 ,2] **& Sébastien BROSSE**[1 ,2]

[1] Université de Toulouse, UPS, ENFA; UMR5174 EDB (Laboratoire Évolution et Diversité Biologique); 118 route de Narbonne, F-31062 Toulouse, France.

[2] CNRS; UMR5174 EDB, F-31062 Toulouse, France.

Corresponding author: Christine Lauzeral, Laboratoire Evolution et Diversité Biologique, U.M.R 5174, C.N.R.S - Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse, France. Email: christine.lauzeral@univ-tlse3.fr

Abstract

Species distribution models (SDMs) link species occurrences to environmental descriptors using species and environmental data that are often recorded at different grain sizes. The upscaling process implied by grain size matching between species data and environmental data may affect the observed species distribution and thus might also modify the SDM-derived species distribution. Here we used four simulated species with different prevalences to determine the effects of grain size on SDM-derived distribution area. We showed that the increase of SDM-derived distribution area with grain size is mainly due to the geometric increase of the area of the observed distribution range used to build the SDMs. Models built using the raw ratio of presences to absences in the learning data set and the maximization of TSS or Kappa as cut-off threshold accurately predicted the observed area whatever the species prevalence and grain size. In addition it should be noted that the commonly used quality indices (AUC, TSS and Kappa) cannot be used to evaluate the accuracy of SDM-derived distribution areas. The grain size of the data used to feed SDMs has to be chosen carefully, depending on the data quality and the goals of the study.

Introduction

The grain size of the data used to feed species distribution models (SDMs) varies with the environmental and climate variables considered as well as with the source of occurrence data, the extent, and the location of the region considered (Guisan and Thuiller 2005). Previous studies have focused on the effects of grain size on model prediction performance(e.g., Guisan et al. 2007), but the influence of grain size on the SDM-derived distribution area (i.e. the surface of the predicted distribution range) remains a poorly addressed question. Seo et al. (2009) first devoted a paper entirely to this question and showed that the SDM-derived distribution areas of nine tree species could undergo up to a four-fold increase when increasing the grain size from 1x1 km² to above 60x60 km². They however did not identify the sources of this increase. Similarly, Hu and Jiang (2010) modelled the potential distribution of a rare gazelle and noticed a sharper increase in the predicted species distribution area (more than 15-fold increase) when increasing the grain size from 1 x 1 km² to 32 x 32 km² despite a slight decrease in model quality through upscaling. The consistency of these trends for various organisms including both plants (Seo et al. 2009) and animals (Hu and Jiang 2010) suggests that the species distribution area increase through grain size increase could have its root in methodological sources.

As increasing the operational grain size obviously induces a geometric increase of the observed species distribution area (if a species is present in only one out of four adjacent cells, merging these four cells when upscaling will lead to a four-fold area increase), it also affects the SDM-derived species distribution area. Nevertheless, the relationship between observed species distribution area and SDM-predicted area through upscaling remains to be clarified. In addition, the geometrical area increase might be sensitive to the species distribution range geometry, explaining why Seo *et al.* (2009) found that the predicted species distribution area

increase depended on the species prevalence. Understanding the sources of variation of SDM-derived distribution areas through upscaling is thus a crucial issue.

Here, we tested how increasing the grain size affects observed species distribution areas and in turn SDM-predicted areas. Virtual species with known prevalence were created to measure how the observed and SDM-derived distribution areas vary through the upscaling process. We first assessed how the geometry of the distribution range affected the observed species distribution area increase through upscaling. We then determined how the SDM-derived distribution areas were linked to the observed areas and how this relationship was affected by species prevalence, presence-absence ratio in the learning data set, cut-off threshold choice and grain size. Finally, we investigated how grain size affected model accuracy using model quality indices commonly found in the literature.

Material and methods

Predictor variables

Eight climate variables were extracted over France from the 30" × 30" resolution WorldClim layers for the period 1961-1990 (Hijmans et al. 2005): precipitation in the driest quarter of the year and in the wettest quarter; average monthly precipitation and precipitation seasonality; mean temperature of the coldest quarter and of the warmest quarter, annual mean temperature and temperature seasonality. These variables were chosen as they are related to the ecological requirements of numerous species, and have often been used in SDMs (Buisson et al. 2010, Marini et al. 2009, Thuiller et al. 2005).

Virtual ecological niches

The virtual species distributions were delineated by constructing two synthetic climate variables (Jimenez-Valverde and Lobo 2007). A normalized principal component analysis

(PCA) was computed on the eight climate variables and the first two axes of the PCA, that accounted for 80% of the total variance, were kept as synthetic variables. In the two-dimensional space created by these two orthogonal axes summarizing climatic variation across France, the virtual species niches were defined as discs centred on coordinates (0,0). All geographic cells falling within the disc for the pair of climate variables were considered as the observed distribution range of the virtual species in France. Using four different disc radiuses, four virtual species were created, with prevalence of 1%, 5%, 15% and 30%, respectively (Fig 1), so as to cover a prevalence ranges similar to those of Seo et al. (2009) and Hu and Jiang (2010).

Grain sizes

Seven grain sizes were selected from 30" × 30" to 32' × 32' (approximately from 1x1 km² to 60x60 km²) (Hu and Jiang 2010, Seo et al. 2009). Species presence or absence and climate variables were upscaled into these grids. Presence per cell was assigned when one or more presence records were found in the merged cells. Predictor variables were upscaled using the mean value of the merged predictor variables.

Models

We used an ensemble modelling approach (Araújo and New 2007). Five classical SDMs (i.e., generalized linear models (GLM); generalized additive models; classification and regression trees; discriminant factorial analysis (DFA) and Random Forest) (e.g., Buisson et al. 2010) were run for each species, at each grain size, using the 8 climate variables. For GLMs and DFAs, squared variables were included in the model to deal with non-linearity. The probability of presence of each species was predicted across the whole region. Then a mean ensemble model was built by averaging the five probabilities of presence and the resulting probability of presence was converted into presence-absence data using three different cut-off thresholds commonly employed (Jimenez-Valverde and Lobo 2007, Liu et al. 2005): (1) by

maximizing the percentage of presences and absences correctly predicted i.e., by maximizing the TSS, (2) by maximizing the Kappa, (3) by using the prevalence in the learning data set.

Datasets

Among the 912730 30”x30” cells covering the area of France, 5000 cells were randomly sampled and then upscaled into all the coarser grains. These 5000 cells were considered as the sampling sites. This operation was repeated 100 times, giving rise to 100 data sets at each of the 7 grain sizes. As a presence-absence ratio of 1:1 has been recommended in some studies whereas others used the raw presence-absence ratio (e.g. Buisson et al. 2008, Jimenez-Valverde and Lobo 2006), both were considered in this study. In the raw presence-absence ratio all the cells were kept whatever the absence-to-presence ratio. In the fixed 1:1 presence-absence ratio, all the presence cells were kept and absence cells were selected randomly from the remaining sampling cells. When insufficient absence cells were available, all absence cells were kept and presence cells were randomly selected to respect the presence-absence ratio of 1:1. This gave rise to a total of 5600 distribution datasets (100 sampling datasets x 4 species x 7 grain sizes x 2 presence-absence ratio). Each of these 5600 datasets was randomly split into two parts: two-thirds of the data were used to calibrate the SDMs and the remaining third was used as a test set. We thus obtained 600 (100 data sets, 2 presence-absence ratios in the learning and test sets, 3 cut-off thresholds) predicted niches at each grain size and for each species.

Effect of grain size on observed area

All the range areas were evaluated as the number of occupied 30”x30” pixels included in the considered range. We first evaluated the observed distribution areas on each of the seven grain sizes and verified that observed areas increased exponentially with grain size. We then tested in which way this increase from one grain size to the following was linked to the geometry of the observed niche. For each species and each grain size (except the coarsest

one), we calculated the mean number of empty cells (N_{emp}) adjacent to the occupied cells. We then calculated the species distribution area increase (A_{inc}) as the ratio between the distribution areas measured for two successive grains and assessed the relationship between N_{emp} and A_{inc} to test for a niche geometry effect on the observed distribution area.

Effect of grain size on SDM-derived distribution area

Regarding SDMs outputs, we first tested the model's ability to predict the observed area. We evaluated the correlation between observed and SDM-derived distribution areas. The ratio between observed and SDM-derived distribution areas was then calculated for each grain size. This enabled us to view the effect of grain size on the accuracy of SDM-predicted distribution area after removing the geometric increase of the observed area. We assessed the results for each species, each presence-absence ratio in the learning data set and each cut-off threshold. Then, for each species, we used a linear model to partition out the variability in the ratio between observed and predicted areas due to each of the other parameters (grain size, presence-absence ratio and cut-off threshold) by using the ratio between the variance explained by one factor and the total variance.

Model accuracy

To assess spatial congruence between the observed and the predicted niche, model accuracy was evaluated using three commonly used metrics: the area under the ROC curve (AUC), a threshold-independent measure; Kappa and TSS, that are two threshold-dependent measures. We also plotted a map of omission and commission errors. For each species, we counted, over the 100 models built using the 100 different learning data sets, the percentage of mispredicting models in each pixel. This was done at each grain size.

Results

Upscaling the grain size from 30”x30” to 32’x32’ caused a sharp increase in observed species distribution areas (a 2.5-fold increase for the common species, and up to 25-fold for the rarest species) (Fig. 2 A). The species distribution area increased exponentially with grain size (Pearson correlation, $r=0.99$, $p<0.001$). Through upscaling, the increase of species distribution areas (log-transformed) was significantly linked to the geometry of the niche (Fig. 2 B) as it was strongly positively correlated to the number of empty cells around occupied cells (Pearson correlation $r=0.99$, $p<0.001$). The increase of species distribution areas was therefore more pronounced for rare species, which distribution range was more fragmented (2.8 empty neighbours for the rarest species vs. 0.6 for the common species at the finest grain size, Fig. 2 B).

Most of the SDM-derived distribution areas increased exponentially with grain size. The Pearson correlation coefficients between grain size and the log of the area were all highly significant ($r>0.93$, $p<0.001$) except for the rarest species, using a presence-absence ratio of 1:1 and especially when Kappa maximisation was used as cut-off (Table S1 in Appendix). In all cases, observed and SDM-derived distribution areas were highly correlated (Pearson correlation, $r >0.94$, $p<0.01$).

Whatever the presence-absence ratio and cut-off threshold considered, SDM predictions of common species (prevalence 15% or 30%) showed a SDM-derived distribution area increase through upscaling ranging from 2 to 3-fold (Fig. 3, Fig. S1 in Appendix). The ratio between observed and SDM-derived distribution area varied slightly but the variance was mostly explained by grain size (Table 1) as the ratio globally decreased during the upscaling process. More precisely, all SDMs accurately predicted the common species distribution areas, as they predicted a species distribution area which was on average 1.11 (± 0.14) times larger than the observed one measured on the same grain size (Fig. 3, Fig. S1 in Appendix).

With regard to rare species (prevalence 1% or 5%), SDM predictions were strongly affected by the presence-absence ratio (raw or 1:1) and the cut-off threshold (Kappa, TSS or prevalence) (Table 1). Using a 1:1 presence-absence ratio, SDM-derived distribution showed an area increase through upscaling ranging from 2 to 4-fold (Fig. S1 in Appendix). These models strongly overpredicted the rare species distribution areas (except at very large grain size). They predicted a species distribution area 6.33 (± 5.89) times larger than the observed one considered on the same grain size (on average over grain sizes) and 18.18 (± 7.65) at the smallest grain size for the rarest species (Fig. S1 in Appendix).

Using a raw presence-absence ratio, the increase of SDM-derived distribution area for rare species through upscaling was strongly influenced by cut-off threshold selection (Fig. 3). The prevalence cut-off provided results similar to those obtained using 1:1 presence-absence ratio. The SDM-derived distribution area showed an increase through upscaling of around 2.5. Ratio analysis showed that the SDM-derived distribution area was on average 4.32 (± 3.92) times larger (15.27 (± 1.3) times larger at the smallest grain size for the rarest species). On the contrary, the models using Kappa and TSS cut-off as thresholds produced SDM-derived distribution area increases through upscaling similar to those observed for the observed niche (Fig. 3). The two cut-off threshold models accurately predicted all species distribution areas, as they predicted species distribution areas which were on average 1.03 (± 0.09) times larger than the observed ones measured on the same grain size (Fig. 3).

Concerning model predictive accuracy, the AUC slightly decreased with grain size (Fig. 4, Fig. S2 in Appendix). This decrease was greater for rare species and for large grain sizes. The pattern did not differ with presence-absence ratio in the learning and test sets. A similar tendency was recorded for Kappa, TSS, sensitivity and specificity (Fig. 5, Fig. S3 in Appendix). Only the Kappa of rare species models built using raw presence-absence ratio and

prevalence as threshold and the sensitivity of rare species models using raw presence-absence ratio and Kappa or TSS as threshold increased with grain size.

Although quality indices were slightly affected by upscaling, the geographical distribution of model omission errors of rare species dramatically varied through the upscaling process. At small grain size, models predicting the rare species distribution often omitted a large part of the geographical distribution while omissions errors of models built at larger grain sizes were more uniformly distributed (Fig 6, Fig. S4 in Appendix). On the contrary, model omission errors were mainly located at the edge of the observed distribution, whatever the grain size.

Discussion

Work focusing on grain size effects has provided mixed results: although Guisan et al (2007) showed that SDM performance was not greatly affected by a 10-fold change in grain size, Seo et al. (2009) demonstrated that changing grain size dramatically increased the SDM-derived species distribution area.

As reported by Seo et al. (2009) using real species, in the present study we observed that the SDM-derived distribution areas of four virtual species increased exponentially with grain size. This increase was primarily due to the increase of the observed distribution area caused by coarsening grain size, mostly at the edge of the distribution due to the presence of the species in cells adjacent to empty cells, revealing a strong geometric effect. Moreover, as geometrical constraints explained most of the observed distribution area increase, its magnitude was strongly influenced by species rarity, as the niches of rare species were much more fragmented than those of more common ones. This trend might be amplified in the case of real species with highly fragmented distributions. Such a situation is especially relevant to endangered species, often affected by fragmentation of their distribution range due to habitat destruction or other human disturbances (e.g., Ewers and Didham 2006). Particular caution

should therefore be taken when choosing grain size to measure distribution areas of species whose populations are fragmented.

Apart from the geometric effect influencing the observed area and hence the SDMs outputs, other parameters can affect the predicted species distribution areas, i.e. the species prevalence, the presence-absence ratio in the learning data set and the cut-off threshold.

The distribution areas of species having a prevalence higher than 10% were accurately predicted whatever the presence-absence ratio in the learning and test set and the threshold used. This contrasts with rare species for which area prediction quality depended on both the presence-absence ratio and the cut-off threshold selection. For these species, a strong area overprediction occurred in 4 out of the 6 presence-absence ratio and cut-off combinations, paralleling therefore the overpredicted prevalence observed by Manel et al. (2001) on rare invertebrate species.

Considering presence-absence ratio for rare species, the raw ratio preserved a better representation of the distribution area, and should hence be preferred to a fixed 1:1 ratio. Indeed, resampling in order to obtain a 1:1 presence-absence ratio yields a loss of information (Jimenez-Valverde and Lobo 2006), especially in unsuitable environmental conditions, leading to an overprediction of the SDMs-derived distribution area. We hence recommend designing SDM to use raw presence-absence ratios, and either a Kappa or a TSS maximisation as cut-off threshold, as these two indices minimize area overprediction.

Considering model predictive accuracy, the slight decrease of most of the quality indices with grain size is consistent with those reported in previous studies (Guisan et al. 2007, Hu and Jiang 2010). Above all, it emphasises that indices classically used in the evaluation of model accuracy (AUC, Kappa, TSS) cannot be used to evaluate the accuracy of SDM-derived distribution areas.

As Kappa is known to be highly sensitive to prevalence in the test set (Allouche et al. 2006), the choice of a raw presence-absence ratio could lead to smaller values of Kappa at fine grain size. But this dependence reflects the chance corrected nature of Kappa (Santika 2011). In our case, Kappa calculated on models built on raw presence-absence ratio by TSS or Kappa maximization did not vary with grain size although upscaling dramatically increased prevalence in the test set. The Kappa increase was only observed when prevalence was used as a cut-off. It was linked to the decrease of rare species area overprediction. The decrease of the number of predicted presence cells decreased the probability of correct presence prediction by chance. Moreover, the decrease of area overprediction did not affect the TSS measure due to the rarity of the species. Our study thus supports the interest of using Kappa to measure model quality.

Although AUC, TSS and Kappa quality indices slightly decreased with grain size, the sensitivity sharply increased for rare species models using raw presence-absence ratio. This was particularly pronounced when the Kappa maximization was used as cut-off threshold. The Kappa maximisation cut-off threshold that is already known as giving accurate prevalence predictions (Freeman and Moisen 2008), also gives the most accurate area predictions, particularly for rare species. But this is done at the expense of model sensitivity as models predicting rare species omitted a large part of the observed distribution. From a practical point of view, the cut-off threshold has thus to be chosen depending on the study goal. If false absences in distribution predictions have to be restricted, TSS should be preferred. This is for example the case for invasive species studies, where the model aims at identifying potential invasion areas. Indeed from an ecosystem management point of view, overpredicting the invasion area is better than omitting potential invasion sites (Leprieur et al. 2009, Mack et al. 2000). In contrast, Kappa should be preferred when searching to predict

potential prevalence or range area. This could be of interest to design conservation strategies and to set-up protected areas for endangered species.

In accordance with Guisan et al. (2007), we confirmed here that grain size slightly affects model performance measures, but we also demonstrated that, using raw presence-absence ratio and Kappa or TSS maximisation as cut-off threshold, range area predictions were hardly affected once the effect of observed area increase was removed. However, grain size affected the geographic distribution of omission errors of models predicting rare species distribution. Indeed, at fine grain size, SDMs failed to predict presence in large geographical regions. Such a bias no longer occurred at coarser grain size. This might be triggered for real species, especially when projecting future species distributions, since parameters like topography or land cover can increase the uncertainty of projections at fine grain size (Wiens and Bachelet 2009). Our results suggest that an optimal grain size probably does not exist and that it has to be selected depending on the data quality and the goals of the study.

Acknowledgements

EDB is part of the "Laboratoire d'Excellence" (LABEX) entitled TULIP (ANR-10-LABX-41).

This study was supported by the BIOFRESH European project (FP7-ENV-2008).

References

Allouche, O. et al. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). – J. Appl. Ecol. 43: 1223–1232.

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. – Trends Ecol. Evol. 22: 42–47.

Buisson, L. et al. 2008. Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors. – Ecol. Freshw. Fish 17: 244–257.

Buisson, L. et al. 2010. Uncertainty in ensemble forecasting of species distribution. – Global Change Biol. 16: 1145–1157.

Ewers, R. M. and Didham, R. K. 2006. Confounding factors in the detection of species responses to habitat fragmentation. – Biol. Rev. 81: 117–142.

Freeman, E. A. and Moisen, G. G. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. – Ecol. Model. 217: 48–58.

Guisan, A. et al. 2007. Sensitivity of predictive species distribution models to change in grain size. – Divers. Distrib. 13: 332–340.

Guisan, A. and Thuiller, W. 2005. Predicting species distribution: offering more than simple habitat models. – Ecol. Lett. 8: 993–1009.

Hijmans, R. J. et al. 2005. Very high resolution interpolated climate surfaces for global land areas. – Int. J. Climatol. 25: 1965–1978.

Hu, J. and Jiang, Z. 2010. Predicting the potential distribution of the endangered Przewalski's gazelle. – J. Zool. 282: 54–63.

Jimenez-Valverde, A. and Lobo, J. M. 2006. The ghost of unbalanced species distribution data in geographical model predictions. – Divers. Distrib. 12: 521–524.

Jimenez-Valverde, A. and Lobo, J. M. 2007. Threshold criteria for conversion of probability of species presence to either-or presence-absence. –Acta Oecol. 31: 361–369.

Leprieur, F. et al. 2009. Scientific uncertainty and the assessment of risks posed by non-native freshwater fishes. – Fish Fish. 10: 88–97.

Liu, C. R. et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. – Ecography 28: 385–393.

Mack, R. N. et al. 2000. Biotic invasions: Causes, epidemiology, global consequences, and control. – Ecol. Appl. 10: 689–710.

Manel, S. et al. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. – J. Appl. Ecol. 38: 921–931.

Marini, M. A. et al. 2009. Predicted climate-driven bird distribution changes and forecasted conservation conflicts in a Neotropical savanna. – Conserv. Biol. 23: 1558–1567.

Santika, T. 2011. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. – Global Ecol. Biogeogr. 20: 181–192.

Seo, C. et al. 2009. Scale effects in species distribution models: implications for conservation planning under climate change. – Biology Lett. 5: 39–43.

Thuiller, W. et al. 2005. Climate change threats to plant diversity in Europe. – P. Natl. A. Sci. USA 102: 8245–8250.

Wiens, J. A. and Bachelet, D. 2009. Matching the multiple scales of conservation with the multiple scales of climate change. – Conserv. Biol. 24: 51–62.

Table 1:

Analysis of variance of the ratio between observed and SDM-derived range areas. % of variance explained by grain size, presence-absence ratio and cut-off.

	Species prevalence			
	1%	5%	15%	30%
Grain size	41.04	36.78	39.00	50.99
Presence-absence ratio	9.45	5.29	2.67	0.06
Cut-off threshold	7.69	23.31	25.44	17.15
Grain size x Presence-absence ratio	13.40	1.09	0.36	1.73
Grain size x Cut-off threshold	8.29	13.00	8.65	8.60
Presence-absence ratio x Cut-off threshold	4.57	9.74	8.95	1.32

Figure captions

Figure 1: The geographic niches of the four species over France. Species prevalence: 1% (black); 5% (dark grey); 15% (grey); 30% (pale grey). Each niche contains the smaller ones.

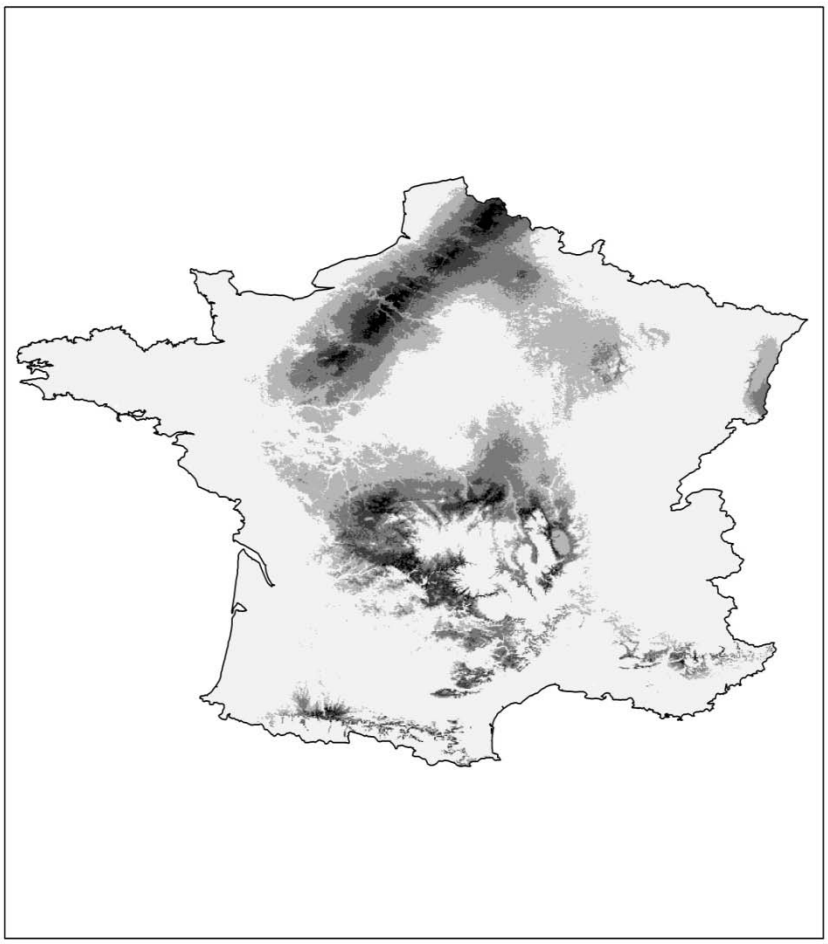


Figure 2: Observed distribution area increase through upscaling. (A) Ratio between the observed distribution area at the 30”x30” grain and the six larger grains. (B) Relationship between the observed distribution area increase (i.e., the ratio between the area measured on two successive grains, log-scaled) and the mean number of empty neighbours adjacent to occupied cells (for the smaller grain). The dashed line represents the linear relationship between the area increase and the number of empty neighbours.

Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

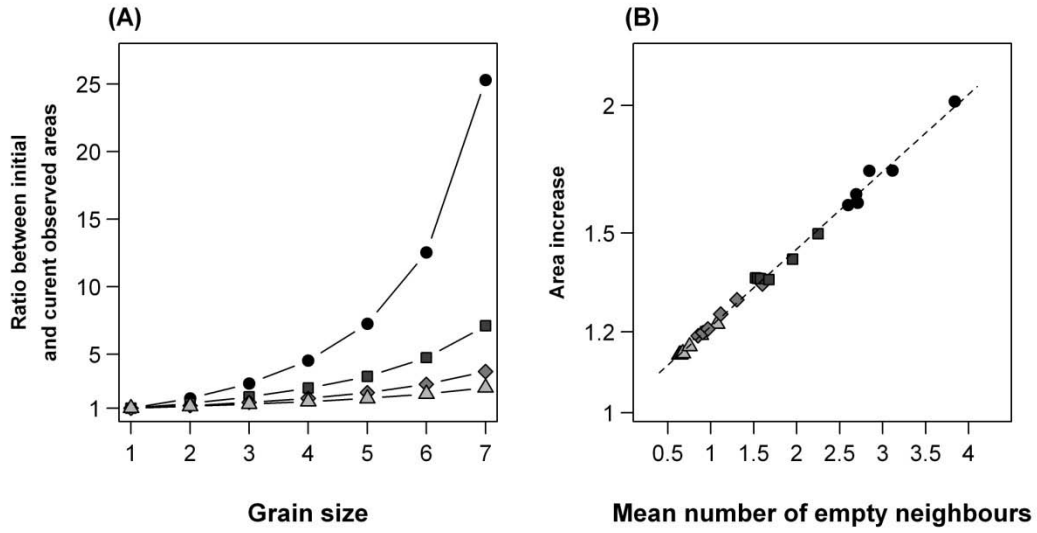


Figure 3: SDM predicted distributions areas through upscaling using raw presence-absence ratio. A, B, C) Mean ratio between the predicted distribution areas at the 30" x 30" grain and the six other grains sizes (log-scaled). D, E, F) Mean ratio between the observed and the SDM-derived distribution areas (log-scaled). In each case, the SDM-derived area was measured on the grain size at which the model was built and compared to the observed area measured on the same grain size. Cut-off thresholds are the maximisation of the TSS (A, D); the maximisation of the Kappa (B, E); the prevalence in the learning data set (C, F). Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

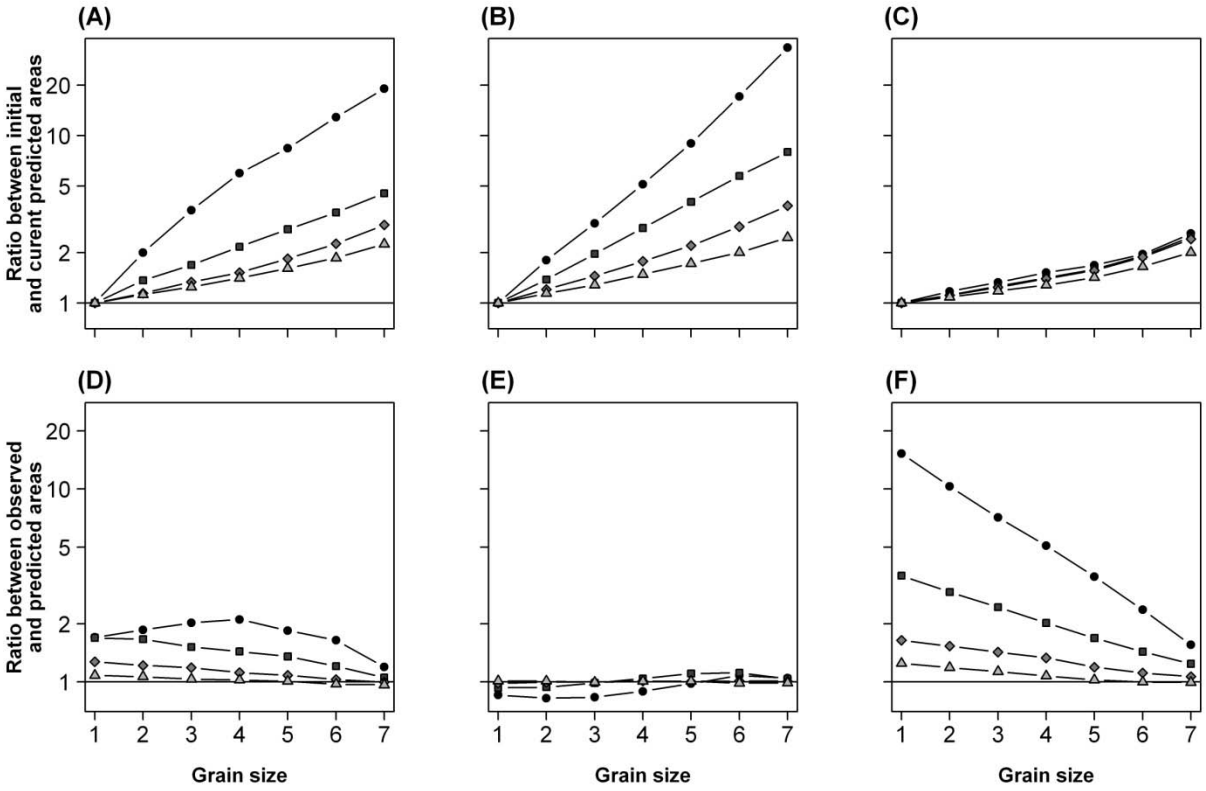


Figure 4: Effect of grain size on AUC (mean value over the 100 test sets) for the four virtual species. All the models were built using raw presence-absence ratio in the learning and test sets. Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

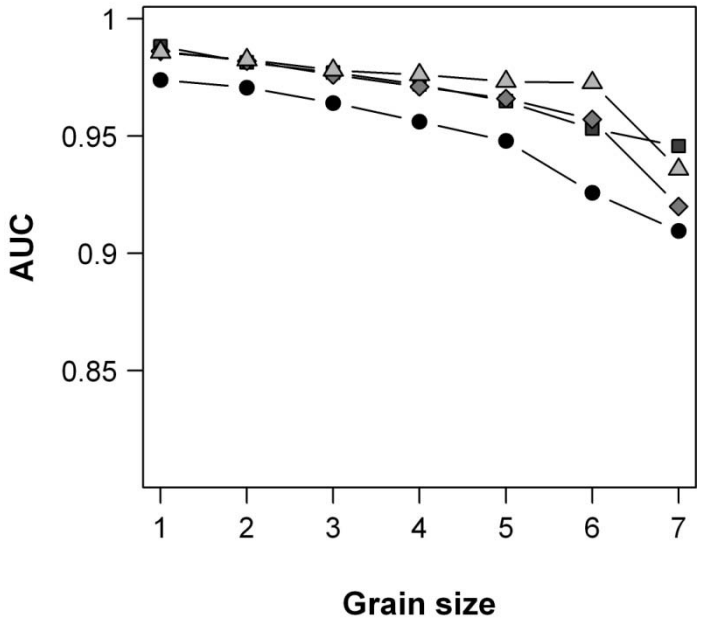


Figure 5: Effect of grain size on model accuracy using Kappa, TSS, sensitivity and specificity (mean value over the 100 test sets) for the four virtual species. All the models were built using raw presence-absence ratio in the learning and test sets. Cut-off thresholds are the maximisation of the TSS (a); the maximisation of the Kappa (b) and the prevalence in the learning set (c). Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

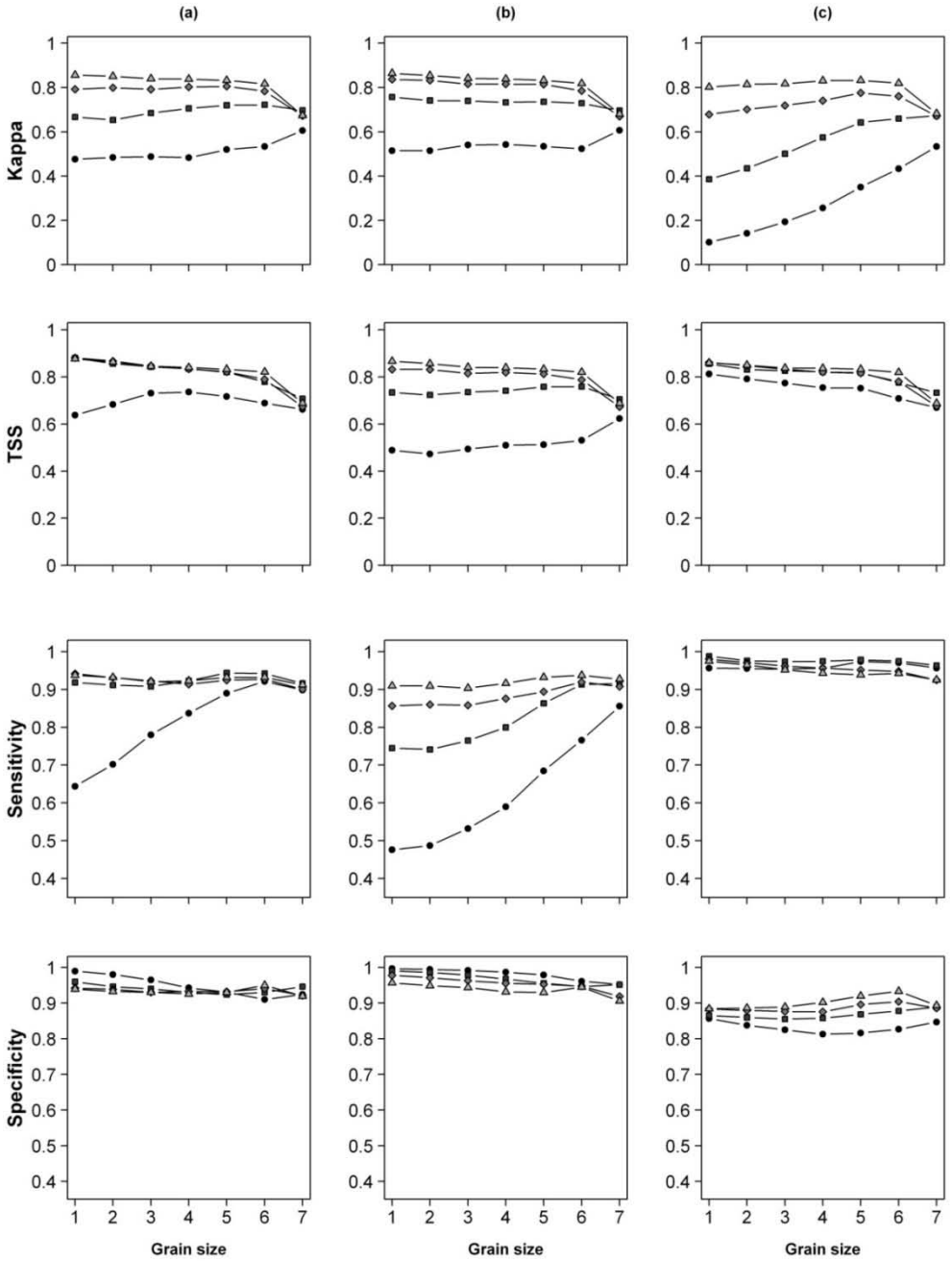
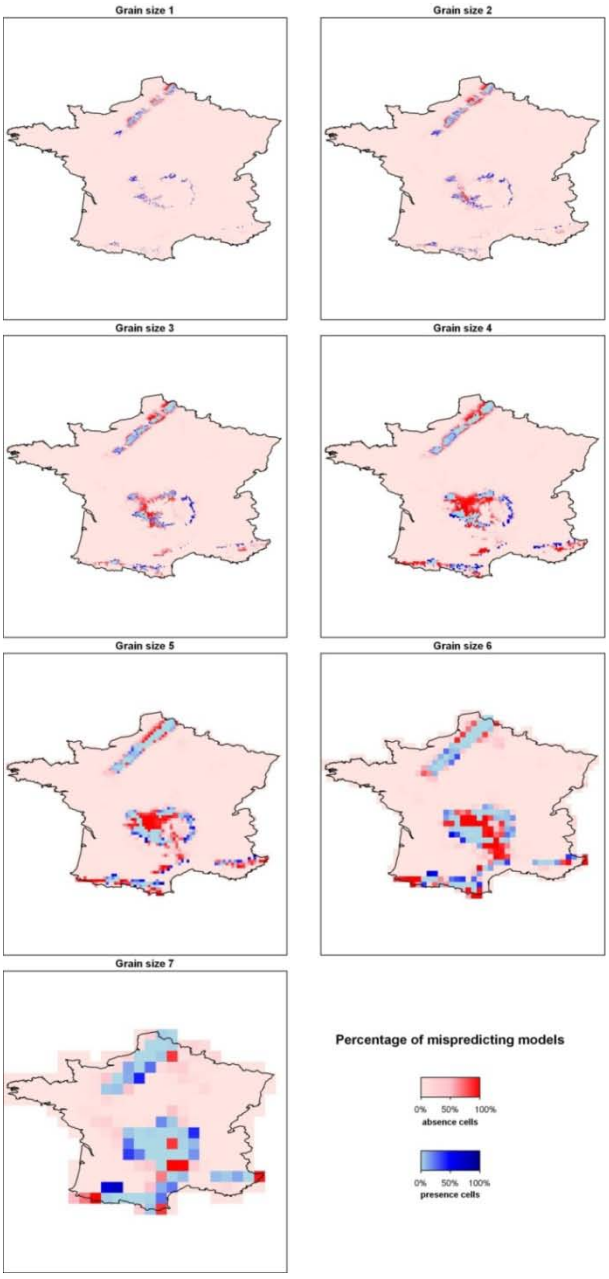


Figure 6: The predicted niches of the rare species (prevalence = 1%) at each grain size. Models were built using raw presence-absence ratio and TSS maximisation as cut-off threshold. The 100 models based on the 100 different learning data sets were used and we evaluated the percentage of mispredicting models in each pixel. The darkest pixels are the most often mispredicted. Note that the model built at the smallest grain omitted a large geographic part of the niche (the south-eastern part).



Appendix

Table S1: Pearson correlation coefficient (r) between the predicted area (log transformed) and the grain size.

	Prevalence of the species	Threshold	Minimal r	Mean r	Maximal r	Number of models with p-value>0.05	Number of models with p-value >0.01
1:1 presence-absence ratio	1%	TSS	0.5420	0.9031	0.9908	5	24
		Kappa	0.1272	0.6435	0.9637	67	88
		Prevalence	0.4485	0.8429	0.9908	19	49
	5%	TSS	0.9289	0.9769	0.9961	0	0
		Kappa	0.9253	0.9730	0.9959	0	0
		Prevalence	0.9417	0.9813	0.9971	0	0
	15%	TSS	0.9492	0.9851	0.9987	0	0
		Kappa	0.9487	0.9836	0.9987	0	0
		Prevalence	0.9663	0.9880	0.9978	0	0
	30%	TSS	0.9689	0.9892	0.9977	0	0
		Kappa	0.9677	0.9883	0.9981	0	0
		Prevalence	0.9686	0.9882	0.9979	0	0
Raw presence-absence ratio	1%	TSS	0.9577	0.9859	0.9985	0	0
		Kappa	0.9622	0.9898	0.9995	0	0
		Prevalence	0.9338	0.9777	0.9990	0	0
	5%	TSS	0.9586	0.9873	0.9986	0	0
		Kappa	0.9826	0.9941	0.9988	0	0
		Prevalence	0.9520	0.9823	0.9974	0	0
	15%	TSS	0.9521	0.9853	0.9987	0	0
		Kappa	0.9815	0.9926	0.9996	0	0
		Prevalence	0.9596	0.9837	0.9973	0	0
	30%	TSS	0.9656	0.9898	0.9992	0	0
		Kappa	0.9805	0.9918	0.9996	0	0
		Prevalence	0.9662	0.9816	0.9958	0	0

Figure S1: SDM predicted distribution areas through upscaling using 1:1 presence-absence ratio. A, B, C) Mean ratio between the predicted distribution areas at the 30" x 30" grain and the six other grains sizes (log-scaled). D, E, F) Mean ratio between the observed and the SDM-derived distribution areas (log-scaled). In each case, the SDM-derived area was measured on the grain size at which the model was built and compared to the observed area measured on the same grain size. Cut-off thresholds are the maximisation of the TSS (A, D); the maximisation of the Kappa (B, E); the prevalence in the learning data set (C, F). Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

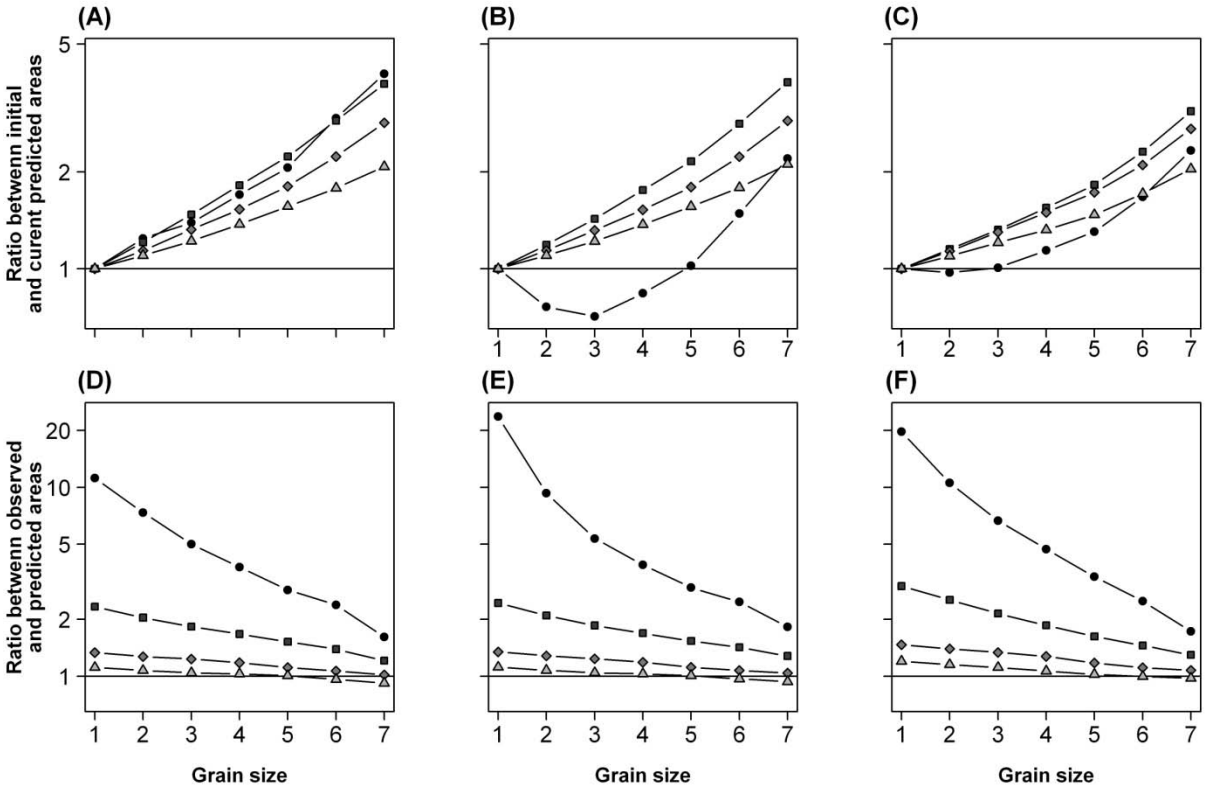


Figure S2: Effect of grain size on AUC (mean value over the 100 test sets) for the four virtual species. All the models were built using 1:1 presence-absence ratio. Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

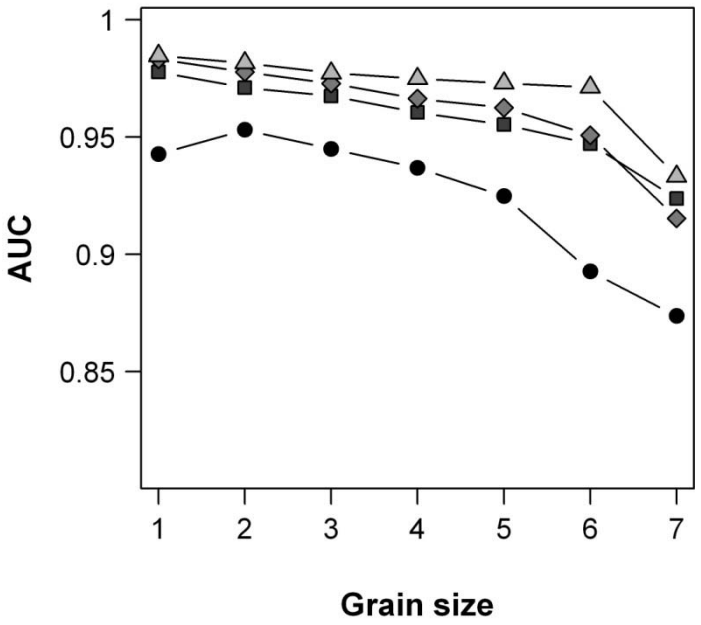


Figure S3: Effect of grain size on model accuracy using Kappa, TSS, sensitivity and specificity (mean value over the 100 test sets) for the four virtual species. All the models were built using 1:1 presence-absence ratio in the learning and test sets. Cut-off thresholds are the maximisation of the TSS (a); the maximisation of the Kappa (b) and the prevalence in the learning set (c). Symbols represent virtual species prevalence. Black circle: 1%; dark grey square: 5%; grey, diamond: 15%; pale grey triangle: 30%.

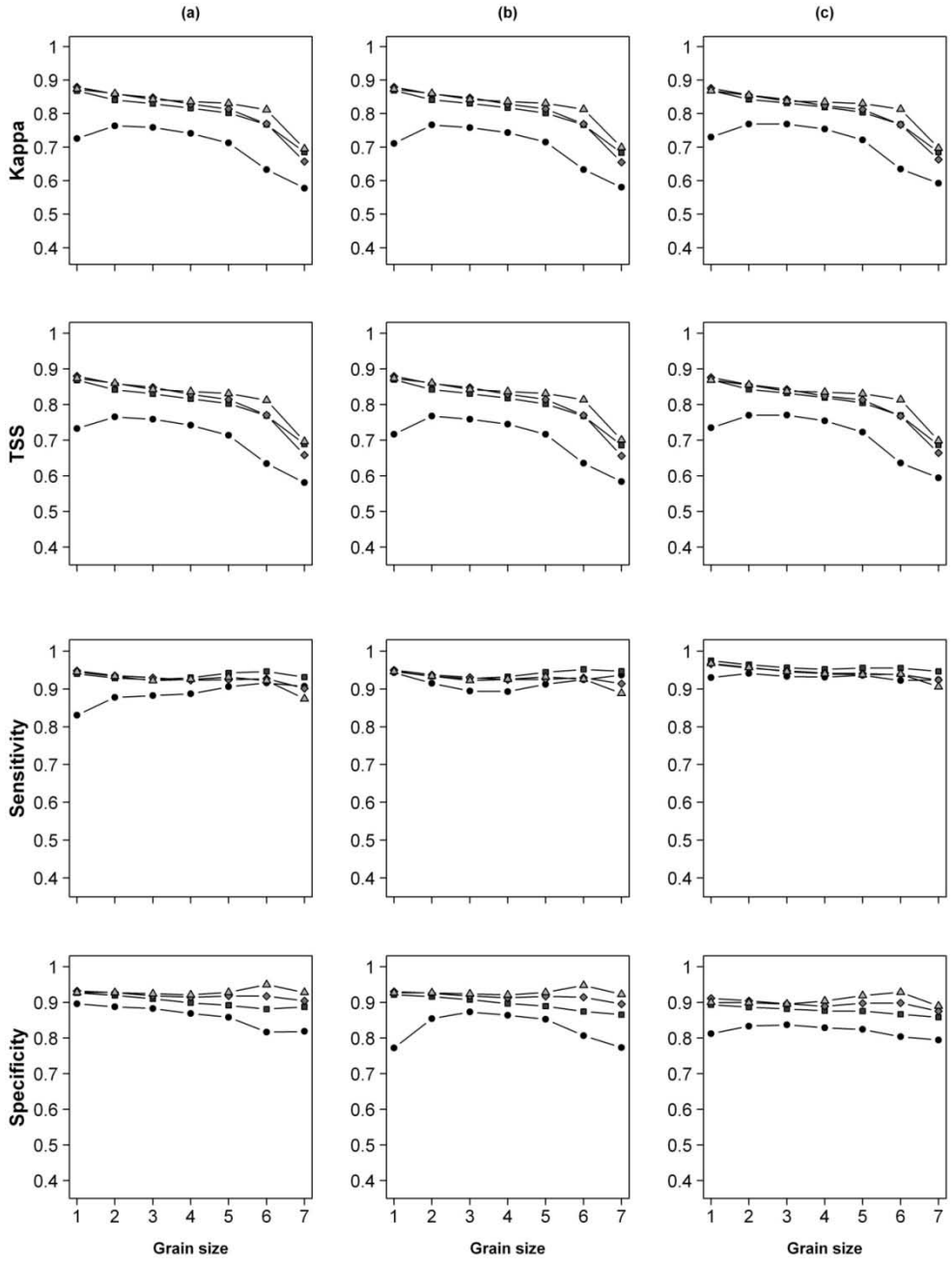
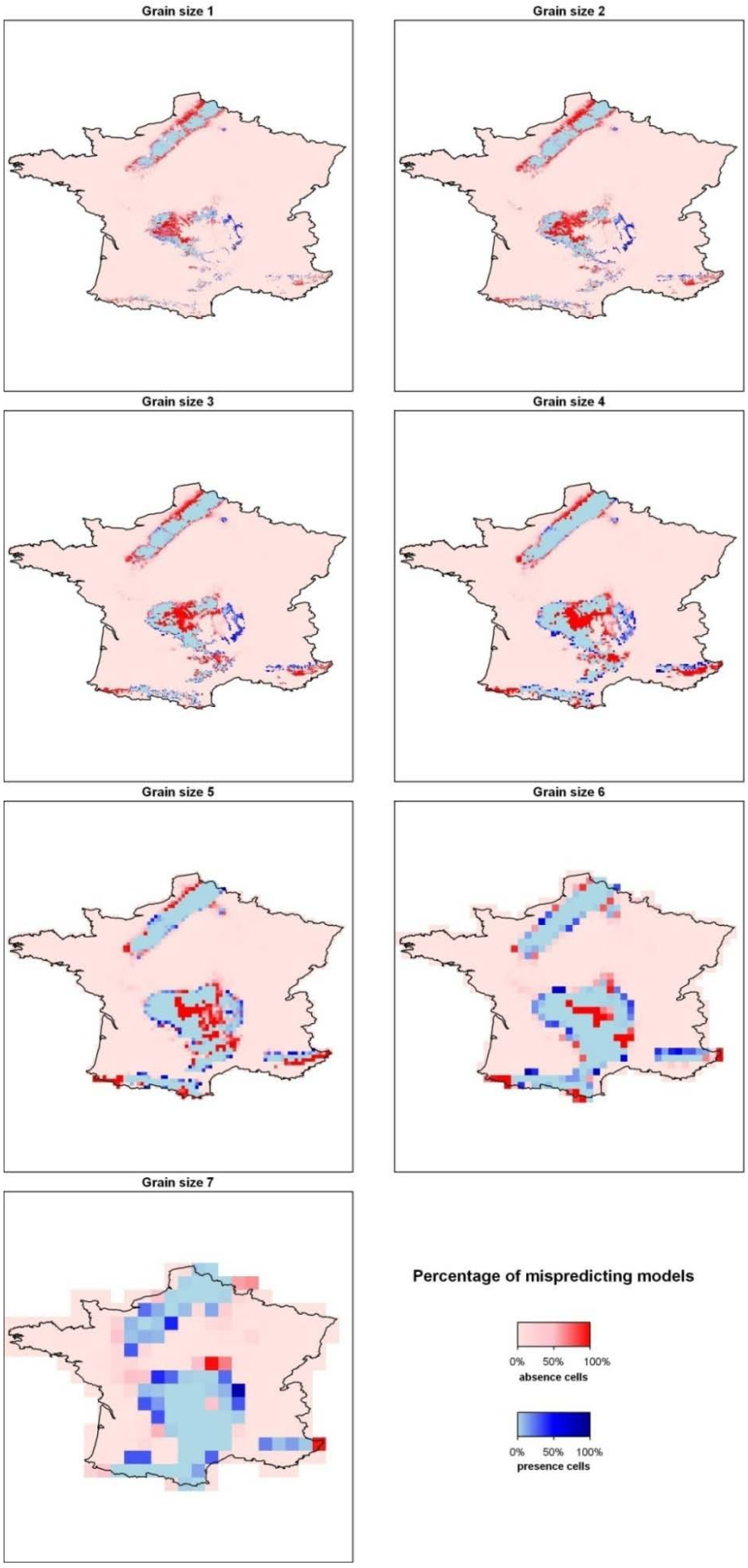
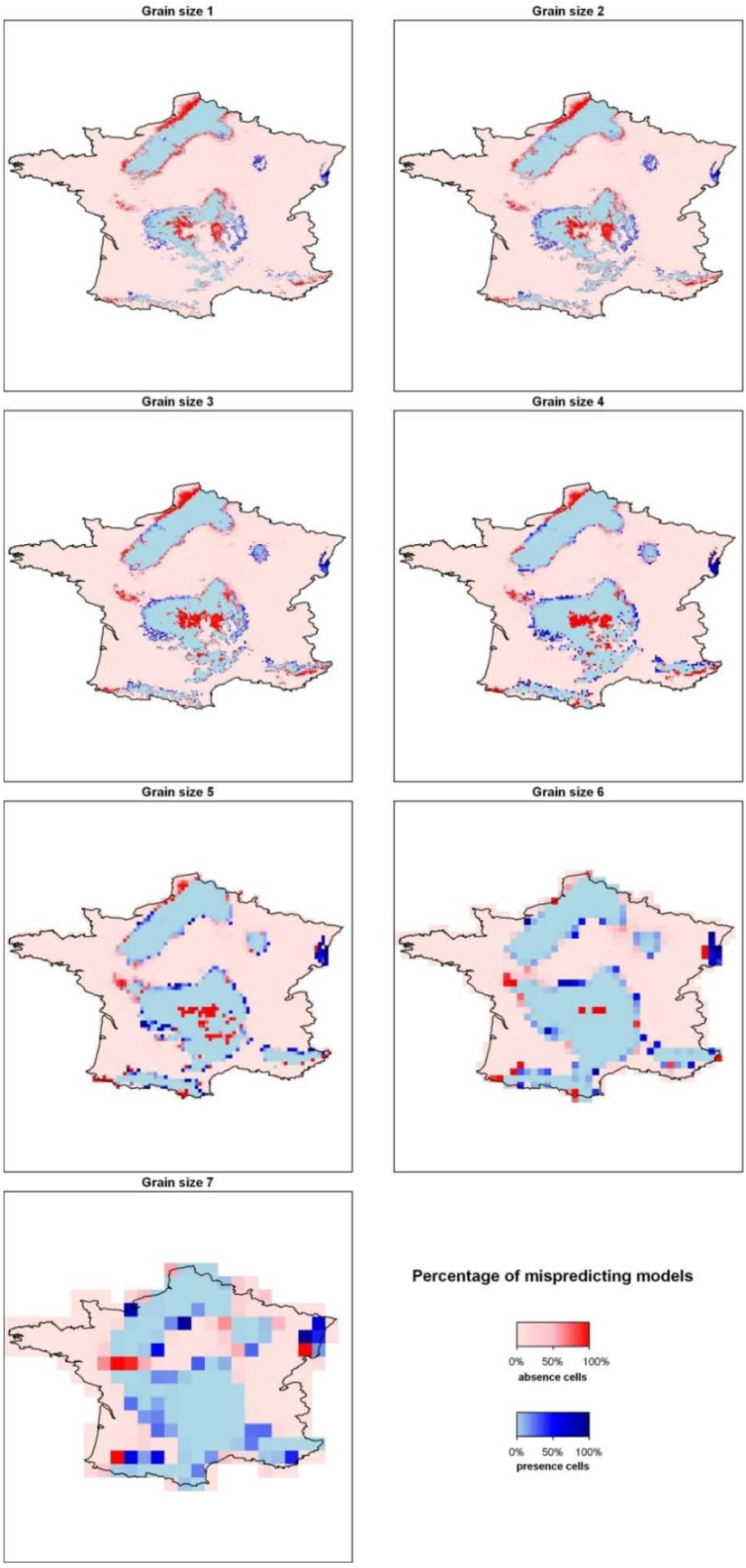


Figure S4: The observed (at the smallest grain size) and predicted (at each grain size) niche for three virtual species with prevalence of 5% (A); 15% (B) and 30% (C). Models were built using raw presence-absence ratio and TSS maximisation as cut-off threshold. The 100 models based on the 100 different learning data sets were used and we evaluated the percentage of mispredicting models in each pixel. The darkest pixels are the most often mispredicted.

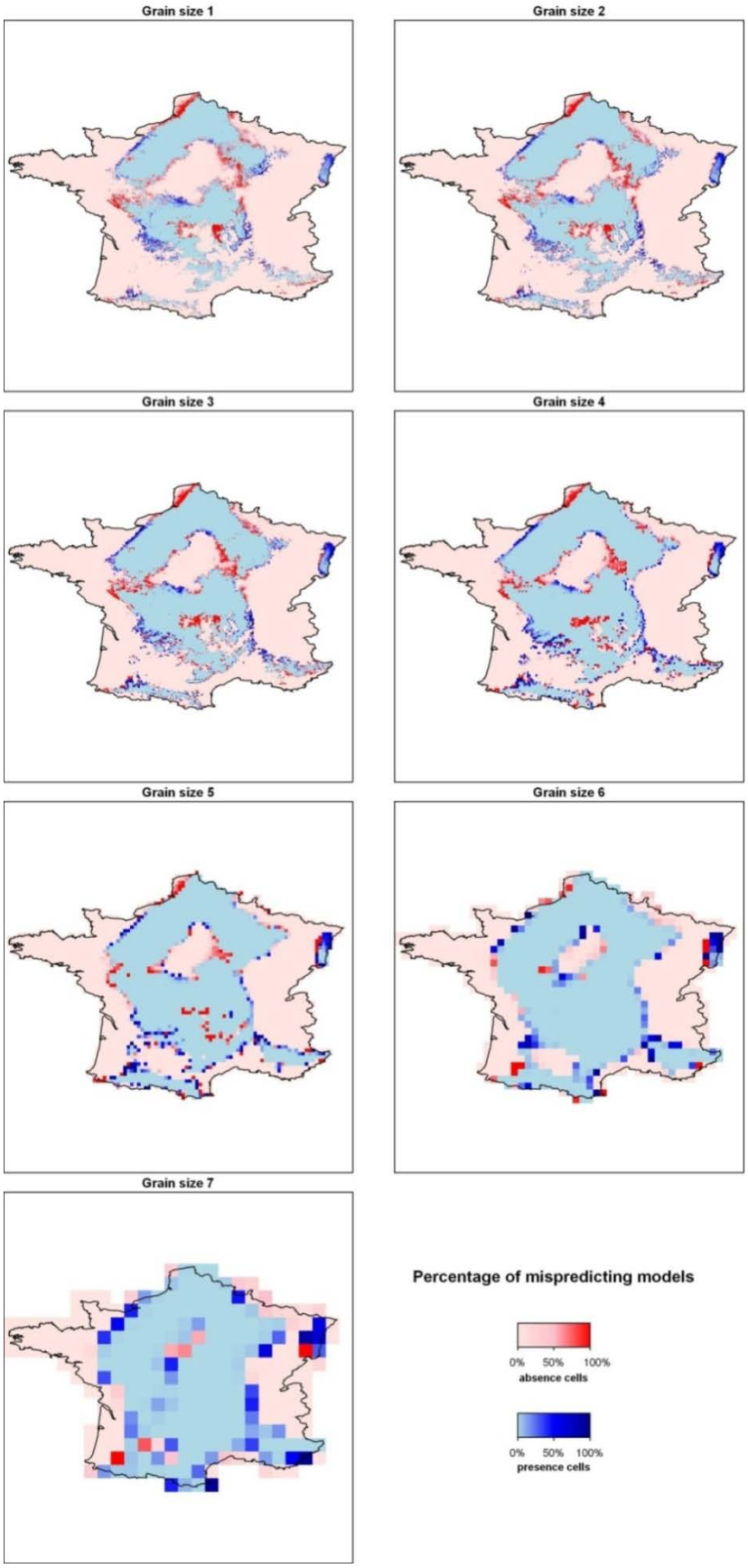
(A)


Grain size 1
Grain size 2
Grain size 3
Grain size 4
Grain size 5
Grain size 6
Grain size 7
Percentage of mispredicting models
0% 50% 100%
absence cells
0% 50% 100%
presence cells

(B)


Grain size 1
Grain size 2
Grain size 3
Grain size 4
Grain size 5
Grain size 6
Grain size 7
Percentage of mispredicting models
0% 50% 100%
absence cells
0% 50% 100%
presence cells

(C)


Grain size 1
Grain size 2
Grain size 3
Grain size 4
Grain size 5
Grain size 6
Grain size 7
Percentage of mispredicting models
0% 50% 100%
absence cells
0% 50% 100%
presence cells

A Journal of Macroecology

Global Ecology and Biogeography

RESEARCH PAPER

Identifying climatic niche shifts using coarse-grained occurrence data: a test with non-native freshwater fish

Christine Lauzeral[1]*†, Fabien Leprieur[2]†, Olivier Beauchard[3], Quiterie Duron[1], Thierry Oberdorff[4] and Sébastien Brosse[1,5]

[1]*Laboratoire Evolution et Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 4, France,* [2]*Laboratoire Ecosystèmes Lagunaires, U.M.R. 5119, CNRS-IFREMER-UM2-IRD- Université Montpellier 2, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France,* [3]*University of Antwerp, Faculty of Sciences, Department of Biology, Ecosystem Management Research Group, Universiteitsplein 1, BE-2610 Antwerpen (Wilrijk), Belgium,* [4]*UMR IRD 207, Biologie des Organismes et des Ecosystèmes Aquatiques, Département Milieux et Peuplements Aquatiques, Muséum National d'Histoire Naturelle, 43 Rue Cuvier, 75231 Paris Cedex, France,* [5]*Laboratoire d'Ecologie Fonctionnelle, UMR 5245, CNRS – Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 4, France*

*Correspondence: Christine Lauzeral, Laboratoire Evolution et Diversité Biologique, UMR 5174, CNRS – Université Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse Cedex 4, France.
E-mail: christine.lauzeral@cict.fr
†Co-first authors.

ABSTRACT

Aim We tested whether coarse-grained occurrence data can be used to detect climatic niche shifts between native and non-native ranges for a set of widely introduced freshwater fishes.

Location World-wide.

Methods We used a global database of freshwater fish occurrences at the river basin scale to identify native and non-native ranges for 18 of the most widely introduced fish species. We also examined climatic conditions within each river basin using fine-grained climate data. We combined this information to test whether climatic niche shifts have occurred between native and non-native ranges. We defined climatic niche shifts as instances where the ranges of a climatic variable within native and non-native basins exhibit zero overlap.

Results We detected at least one climatic niche shift for each of the 18 studied species. However, we did not detect common patterns in the thermal preference or biogeographic origin of the non-native fish, hence suggesting a species-specific response.

Main conclusions Coarse-grained occurrence data can be used to detect climatic niche shifts. They also enable the identification of the species experiencing niche shifts, although the mechanisms responsible for these shifts (e.g. local adaptation, dispersal limitation or physiological constraints) have yet to be determined. Furthermore, the coarse-grained approach, which highlights regions where climatic niche shifts have occurred, can be used to select specific river basins for more detailed, fine-grained studies.

Keywords

Bioclimatic models, climate mismatch, freshwater fish, invasion, risk assessment, river basins.

INTRODUCTION

Bioclimatic models of species distributions are increasingly being used to predict the establishment and spread of non-native species over new areas, and to forecast range shifts in invasive species due to climate change (e.g. Thuiller *et al.*, 2005; Jeschke & Strayer, 2008; Britton *et al.*, 2010). These models are built under the assumption that species are in equilibrium with the climatic conditions encountered in their native ranges (i.e. their realized niche; Hutchinson, 1957) and that they tend to maintain ancestral ecological requirements in their non-native range (i.e. niche conservatism; see Jeschke & Strayer, 2008). Under these assumptions, the climate range where a species can become established can be predicted by fitting models with climate data from its native range (i.e. climate matching). This climate or environmental matching approach has been widely applied in invasion risk assessment (e.g. Bomford *et al.*, 2009).

Such approaches have recently been criticized (e.g. Broennimann & Guisan, 2008) because the spatial distribution of a species is not only constrained by current climate but also by historical and biotic factors such as barriers to dispersion, biotic interactions and stochastic events (Jiménez-Valverde *et al.*, 2008). Consequently, a number of studies have shown that when models are trained (i.e. parameterized) using data from the

DOI: 10.1111/j.1466-8238.2010.00611.x
www.blackwellpublishing.com/geb

native range, they tend to underpredict the non-native range, i.e. models were unable to predict the full extent of invasion (e.g. Broennimann *et al.*, 2007; Loo *et al.*, 2007; Medley, 2010). The reason is that species might be able to establish and spread into localities (or regions) that are climatically distinct from those encountered within the native range (i.e. a climatic niche shift).

Instances of climatic niche shifts have recently been reported for a wide range of plants and animals (e.g. Fitzpatrick *et al.*, 2008; Rödder & Lötters, 2009; Medley, 2010). These approaches typically use fine-grained data (e.g. 0.1° to 0.5° latitude and longitude grid cells), hence requiring the assembly of numerous local occurrence data in both the native and the non-native ranges. However, such detailed information is rarely available on a large scale (Pyšek *et al.*, 2008), limiting the identification of climatic niche shift to a restricted number of well-studied species. In contrast, much more information is available at a coarser spatial grain (e.g. ecoregion, country, province) through the use of natural history atlases or regional biodiversity assessments (e.g. DAISIE, 2009). Developing methods to identify niche shifts using these coarse-grained occurrence data would therefore considerably increase the pool of species for which a climatic niche shift can be identified.

We used a global database of freshwater fish to test whether coarse-grained occurrence data can be used to detect climatic niche shifts between native and non-native ranges. Specifically, we used river basins as our sampling unit (see Leprieur *et al.*, 2008). Freshwater fish distributions are influenced by many factors operating at different spatial scales (reviewed in Jackson *et al.*, 2001). Large scale (e.g. among river basins) present-day patterns of freshwater fish distribution are influenced by historical connections between river basins, Earth history events (e.g. Quaternary glaciations, orographic formation) and environmental constraints (e.g. climatic zones, biomes) (e.g. Jackson & Harvey, 1989; Leprieur *et al.*, 2009a). Smaller-scale patterns (e.g. within a given river basin) of fish distribution are mainly influenced by geometry of the river network and a combination of abiotic and biotic factors, including temperature and hydrology (Jackson *et al.*, 2001). In the present study, we analysed patterns of fish distribution among river basins for 18 introduced species that are known to be widely established beyond their native range.

MATERIALS AND METHODS

The use of coarse-grained occurrence data (i.e. river basin-scale occurrence data) to detect potential climatic niche shifts can present major limitations. For instance, averaging values of a climatic variable (e.g. annual precipitation) over large and heterogeneous areas may strongly bias the estimation of the climatic niche of a freshwater fish and lead to false claims of niche shifts between native and non-native habitats. To overcome such limitations, we simultaneously analysed coarse-grained fish occurrence data at the river basin scale and fine-grained climate data (i.e. 0.5° × 0.5° gridded climate data that account for the full range of climatic variation within a river basin). Thus, we assumed that a species present in a given river basin can exist anywhere within that basin, and that it is compatible with the full range of climatic or environmental conditions encountered throughout the basin. Climatic niche shifts were then defined as instances where the range of an environmental variable within the native basins exhibited zero overlap with the same variable's range in at least one non-native basin. Notably, this conservative method will increase the probability of Type II error (i.e. failing to detect climatic niche shifts when they have, in fact, occurred), but it is also robust to Type I error (i.e. falsely claiming that climatic niche shifts have occurred).

We used the database of Leprieur *et al.* (2008), which documents occurrences of the world's freshwater fishes at the river basin scale (i.e. complete rivers, from the headwaters to the ocean). Among the 1055 river basins available in our database, the geographic extent of 938 basins dispersed throughout the world was available in a digital format.

For each of the 938 river basins, we collected values of eight climatic variables over the whole surface area from 0.5° × 0.5° gridded climate data (Leemans & Cramer, 1991; New *et al.*, 1999): precipitation in the driest month (P_{min}); precipitation in the wettest month (P_{max}); coefficient of variation of the monthly precipitation (P_{cv}); number of rainy days (N_{rd}); mean temperature of the coldest month (T_{min}); mean temperature of the warmest month (T_{max}); coefficient of variation of mean monthly temperature (T_{cv}); and mean annual temperature range (T_{ampl}). These climatic variables are often used in broad-scale studies of freshwater fish distributions (e.g. Minns & Moore, 1995; Chu *et al.*, 2005; Leprieur *et al.*, 2009a), because broad-scale physiological and ecological requirements of freshwater fish species are largely related to temperature and hydrology (Matthews, 1998). Moreover, the Pearson correlations between the eight variables remain low. Although all of the 28 determination coefficients were significant ($P < 0.05$), all were lower than 0.65, except $T_{min} - T_{cv}$ (see Table S1). However, we kept these two variables to maintain a similar approach for temperature and precipitation patterns. For the same reason, overall mean values of temperature and precipitation were not used as they were highly redundant with T_{min} [Pearson determination coefficient r^2 ($T_{min} - T_{mean}$) = 0.96; $P < 0.001$] and P_{max} [Pearson determination coefficient r^2 ($P_{max} - P_{mean}$) = 0.81; $P < 0.001$]. Air temperatures were used as a substitute for water temperatures, which are not currently available for many river basins. This is generally acceptable because streams and rivers are well-mixed water bodies that readily exchange heat with the atmosphere, and it has been empirically demonstrated that air and river water temperatures are strongly positively correlated (e.g. Caissie, 2006).

We then selected fish species that had been widely introduced outside their native range (Lever, 1996) and that are native to more than 15 river basins to ensure that the native basins available in our dataset are representative of a substantial part of the native range. *Gambusia affinis* and *Gambusia holbrooki* were considered together due to their uncertain taxonomic status in the literature and to their similar ecological requirements (Pyke, 2008). For the 18 resulting fish species (see Table 1), we gathered species occurrences per river basin and distinguished between

Table 1 Native and non-native species occurrences (i.e. number of basins) and percentage of non-native basins experiencing a niche shift for at least one climate variable.

Species	Native occurrences	Non-native occurrences	Percentage of non-native basins with niche shift
Ameiurus melas	18	49	14.3
Carassius auratus	49	164	12.8
Carassius carassius	30	54	55.6
Cyprinus carpio	34	245	20.4
Gambusia sp.	52	206	20.4
Ictalurus punctatus	33	44	25.0
Lepomis cyanellus	20	46	23.9
Lepomis gibbosus	28	52	23.1
Lepomis macrochirus	44	63	19.0
Micropterus salmoides	48	100	24.0
Oncorhynchus mykiss	56	189	27.5
Perca fluviatilis	103	64	42.2
Pseudorasbora parva	29	45	15.6
Salmo trutta	141	131	21.4
Salvelinus fontinalis	66	71	40.8
Sander lucioperca	32	48	43.8
Thymallus thymallus	35	33	12.1
Tinca tinca	79	46	10.9

native and non-native occurrences. A species was considered non-native when: (1) it did not historically occur in a given basin, and (2) it was successfully established (i.e. had self-reproducing populations) (see Leprieur *et al.*, 2008; Blanchet *et al.*, 2009). For a few basins (no more than four per species) the native or non-native status was uncertain in the literature, and the species was considered as native in these basins to avoid identifying undue niche shifts.

We assumed that if a species is present (native or non-native) in a river basin, it is potentially present in all of the 0.5° × 0.5° grid cells encompassed by the river basin. We then defined climatic niche shifts as instances where the climate characteristics of one or more non-native basins exhibited zero overlap with climate conditions observed throughout the native range. As a climatic niche shift can result from either an increase or a decrease in a climatic variable, we distinguished between positive and negative shifts. For each species and for each climatic variable, we determined the percentage of non-native basins for which a positive or negative climatic niche shift was identified. To test for potential bias due to native occurrence sampling of each species we measured the relationship (Pearson's correlation) between the per species percentage of non-native basins experiencing climatic niche shifts and: (1) the number of basins in the native range, and (2) the native range area measured as the number of pixels of 0.5° × 0.5°.

We then tested for a common pattern of climatic niche shifts among species depending on their biogeographic origin (i.e. Nearctic versus Palaearctic) and their thermal guild (i.e. cold water versus cool water; Scott & Crossman, 1973; Keith & Allardi, 2001). We compared the per species percentage of non-native basins experiencing climatic niche shifts between the defined groups using a Mann–Whitney test. For each climate variable, positive and negative shifts were considered as distinct, hence resulting in 10 climate variables (as two variables exhibited both positive and negative shifts).

RESULTS

Climatic niche shifts were identified for each of the 18 studied species (Table 1). Using the eight selected climatic variables, each species experienced a shift between native and non-native ranges for at least one climate variable in, on average, 25% of its non-native basins. There was, however, large variation among species (see Table 1). For instance, the tench (*Tinca tinca*) exhibited climatic niche shifts in about 10% of its non-native basins whereas the crucian carp (*Carassius carassius*) exhibited climatic niche shifts in more than half of its non-native basins. The percentage of non-native basins presenting climatic niche shifts was not significantly correlated with the number of basins in the native range (Pearson correlation $r = 0.063$; $P = 0.81$) nor to the native range area (Pearson correlation $r = 0.329$; $P = 0.18$). Five out of the 18 species exhibited a climatic niche shift for more than half of the eight climate variables (Figs 1 & S1). For instance, the rainbow trout (*Oncorhynchus mykiss*) exhibited a climatic niche shift for each climatic variable (Fig. 1a). The introduction patterns of that species are largely documented (Fausch *et al.*, 2001; Crawford & Muir, 2008), and although its establishment success remains uncertain in some places, it has become established in large areas throughout the world (Fig. 1b). Some of these river basins are characterized by a lower temperature and precipitation variability and by a warmer winter temperature than the basins where that species is native (Fig. 1a).

Considering climatic variables for the overall set of species revealed little consistency among the species: two variables (P_{max} and N_{rd}) exhibited both negative and positive climate shifts,

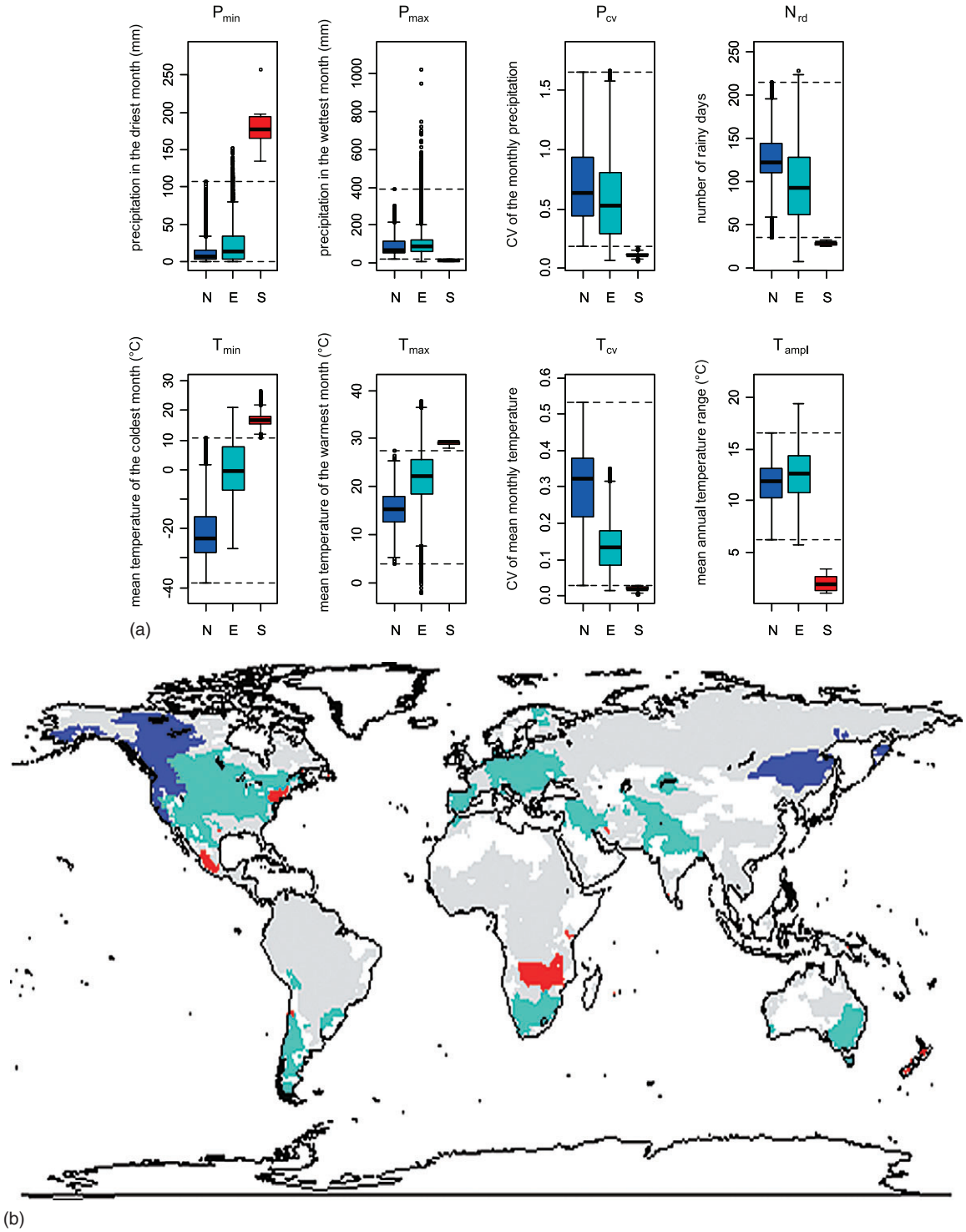


Figure 1 (a) Boxplots representing the climatic range of the rainbow trout *Oncorhynchus mykiss*: For each climate variable, native basins (N, blue), non-native basins without climatic shift (E, turquoise) and non-native basins with climatic shift (S, red) were separated. A climatic niche shift is observed when all the values inside a basin lie outside the two horizontal lines (corresponding to the extreme values inside the native area). The variables used are: precipitation of the driest month (P_{min}); precipitation of the wettest month (P_{max}); coefficient of variation of the monthly precipitation (P_{cv}); number of rainy days (N_{rd}); mean temperature of the coldest month (T_{min}); mean temperature of the warmest month (T_{max}); coefficient of variation of mean monthly temperature (T_{cv}); and mean annual temperature range (T_{ampl}). (b) World-wide distribution of *O. mykiss* based on the 938 basins considered: native basins, non-native basins without climatic shift or non-native basins with climatic shift. Basins available in our database where *O. mykiss* is absent are in grey, areas not covered by our database are in white. Note that the native or non-native status of rainbow trout remains uncertain in the west of the Kamchatka Peninsula basins. To avoid identifying undue niche shift rainbow trout was considered as native in these basins.

Table 2 Environmental variables, percentage of species exhibiting shifts and percentage of basins in which at least one species exhibited a shift.

Environmental variable	Percentage of shifting species	Percentage of exotic shifting basins
P_{min}+	22.2	1.1
P_{max}+	22.2	3.5
P_{max}−	22.2	1.3
N_{rd}+	11.1	1.3
N_{rd}−	66.7	9.7
T_{min}+	38.9	10.4
T_{max}+	5.6	0.2
P_{cv}−	27.8	9.1
T_{cv}−	88.9	26.3
T_{ampl}−	50.0	8.6

The variables used are: precipitation of the driest month (P_{min}); precipitation of the wettest month (P_{max}); coefficient of variation of the monthly precipitation (P_{cv}); number of rainy days (N_{rd}); mean temperature of the coldest month (T_{min}); mean temperature of the warmest month (T_{max}); coefficient of variation of mean monthly temperature (T_{cv}); mean annual temperature range (T_{ampl}). + corresponds to positive shifts, – to negative shifts. The variables experiencing no shift have been removed.

depending on the species. Three other variables (P_{cv}, T_{cv} and T_{ampl}) exhibited only negative shifts, and the three remaining variables (P_{min}, T_{min} and T_{max}) exhibited positive shifts (Table 2). The distribution of shifts also differed: T_{cv} and N_{rd} exhibited at least one shift for more than two-thirds of the species (89% and 67%), whereas T_{cv} exhibited shifts in about one-quarter of the basins and N_{rd} exhibited shifts only in 10% of the basins (Table 2).

Mann–Whitney tests revealed no significant differences between fish species according to their thermal requirements (coldwater versus coolwater species) for any of the 10 climatic variables. With regard to the biogeographic origin of species, Mann–Whitney tests revealed a significant difference between species for only one variable, namely the number of rainy days ($P < 0.01$). Actually, Nearctic species exhibited a much greater number of rainy days shifts than Palaearctic ones. Many of these shifts were located in the south of the United States and in Mexico (i.e. around the native area), except for *Gambusia* sp. which also exhibited numerous shifts in central Asia.

DISCUSSION

For all the species considered in this study, our results showed a niche shift between native and non-native ranges for at least one climatic variable. With these results, the growing literature on climate mismatch for a wide variety of organisms (Broennimann *et al.*, 2007; Fitzpatrick *et al.*, 2008; Beaumont *et al.*, 2009; Rödder & Lötters, 2009; Medley, 2010) is now extended to freshwater fish. The fact that all the considered species are experiencing a niche shift is probably linked to the fact that their realized niches actually don't encompass their entire physiological and ecological ranges (Rosenfield, 2002). Indeed, native species distribution is strongly limited by species incapacity to cross dry land or survive in marine environment (Hugueny, 1989). Our results also show that climatic niche shifts can be detected using coarse-grained data. Importantly, these niche shifts were detected even though we used a highly conservative procedure (i.e. zero overlap in environmental variables between native and non-native basins). It should, however, be noted that our method probably overestimated species climate ranges, as species were assumed to be ubiquitous throughout each river basin. It is therefore highly likely that we underestimated the frequency of climatic niche shifts. It is also possible that sampling artefacts occurred due to incomplete environmental sampling in the native area. This bias is, however, unlikely, because no relationship was found between the surface area (or the number of river basins) in the native range and the percentage of non-native river basins experiencing climatic niche shifts.

Focusing on individual species and specific locations might help to better understand the climatic niche shifts observed. For example, the rainbow trout (and the brown trout) was largely introduced in New Zealand streams and rivers (Townsend, 1996) encountering a more stable climate, which explains a shift toward lower amplitude of temperature and lower coefficients of variation. Here the establishment of trout may have been facilitated by the lack of native enemies, diseases and competitors resulting in a higher tolerance over a wider range of environmental conditions in novel habitats (Moyle & Light, 1996; Townsend, 1996). Moreover, elevated winter temperatures have positive effects on juvenile growth (Morgan *et al.*, 1998) and rainbow trout acquire a higher thermal tolerance for hatching and egg development under a warmer climate (Ineno *et al.*, 2005), which parallels our present temperature mismatch findings: the shift of the rainbow trout to higher minimal temperature in some Mexican and African river basins. From a broader point of view, climate niche shifts have been attributed to three non-mutually exclusive mechanisms: (1) the rapid evolution of species when introduced to novel environments, which may allow them to advance beyond the limits of their climate distribution in their native range (Pearman *et al.*, 2008); (2) physiologically suitable environmental conditions in the non-native

range which are not found in native habitats because of historical or geographical constraints on colonization (Jiménez-Valverde *et al.*, 2008; Leprieur *et al.*, 2009a); and (3) the lack of native predators, diseases and competitors (i.e. enemy release), which can result in higher tolerance to extreme biotic or abiotic conditions (Moyle & Light, 1996; Townsend, 1996). Our coarse-grained data did not allow the relative roles of these three mechanisms to be disentangled, but complementary experimental and fine-grained field studies could help to determine and quantify the mechanisms responsible for climatic niche shifts. Extending these considerations to a multispecies context, there was little similarity between species concerning the climate features to do with climatic niche shift variables, nor in the way they vary. Neither geographical origin nor thermal preferences appear as a strongly significant factor explaining a multispecies response, leaving open the question of the causes of climatic niche shifts (local adaptation, dispersal limitation or physiological tolerance). For instance, we recommend further studies to disentangle the causes of climatic niche shifts focusing on a per species analysis, rather than adopting a multispecies approach.

Overall, our results have important implications for the application of both bioclimatic models and invasion risk assessments. Bioclimatic models are increasingly being used by conservation biologists to forecast the future ranges of both native and non-native species in the face of climate change (Jeschke & Strayer, 2008). Our results suggest, however, that bioclimatic models are likely to underestimate the spread of colonizing species when they are trained or parameterized using environmental data from species native ranges, especially under projected climate change scenarios. Such discrepancies have been recently highlighted for particular plant and invertebrate taxa (i.e. Broennimann *et al.*, 2007; Loo *et al.*, 2007; Fitzpatrick *et al.*, 2008; Medley, 2010). With regard to freshwater fish, we strongly recommend that future studies using bioclimatic models consider climatic conditions found in both native and known non-native ranges so as to consider: (1) a wider sampling of environmental variation (see Menke *et al.*, 2009), and (2) a wider range of climatic conditions in which the species has become established. For instance, this will give a clearer picture of the potential climatic range and hence will reduce uncertainty when assessing the risks posed by non-native freshwater fish (Leprieur *et al.*, 2009b). By the same logic, however, we must caution that such a procedure will be much less informative for potential invaders that have not yet expanded or that have been rarely introduced out of their native range. For such species, bioclimatic models might produce an incomplete picture of their colonization potential (e.g. Loo *et al.*, 2007).

As demonstrated here, coarse-grained occurrence data can be used to identify climatic niche shifts. This is important because a vast number of coarse-grained occurrence data have been published in regional atlases, for a wide variety of taxonomic groups (e.g. DAISIE, 2009). These data can facilitate invasion risk assessments when detailed, local-scale occurrence data are lacking, which is the case for many freshwater fishes but also for most organisms on Earth (Pyšek *et al.*, 2008; Leprieur *et al.*, 2009b). It should, however, be noted that our coarse-scale approach probably underestimates climatic niche shifts. We therefore suggest that it can serve as a first step to identify species experiencing climatic niche shifts, or be used to predict regions where climatic shifts are likely to occur. It might then guide future fine-grained studies in identifying the exact nature and extent of the climatic niche shift observed (see Pearman *et al.*, 2008) and in determining which mechanisms underlie observed patterns (e.g. local adaptation, dispersal limitation or physiological constraints).

To conclude, our study has demonstrated that climatic niche shifts can be identified using coarse-grained species occurrence data. However, the establishment of a species outside its native range is driven by multiple biotic and abiotic factors acting at different spatial resolutions (Lockwood *et al.*, 2007). We thus strongly encourage future studies to extend the present results by applying a multiscale approach. The development of global-scale databases at both fine and coarse spatial resolutions is urgently needed to draw baseline generalities in invasion ecology (Cadotte *et al.*, 2006) which would help managers to prevent future species invasions.

ACKNOWLEDGEMENTS

We thank Pablo Tedesco for his helpful comments on this manuscript. This study was supported by the ANR 'Freshwater Fish Diversity' (ANR-06-BDIV-010, French Ministry of Research) and the BIOFRESH European project (FP7-ENV-2008).

REFERENCES

Beaumont, L.J., Gallagher, R.V., Thuiller, W., Downey, P.O., Leishman, M.R. & Hughes, L. (2009) Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. *Diversity and Distributions*, **15**, 409–420.

Blanchet, S., Leprieur, F., Beauchard, O., Staes, J., Oberdorff, T. & Brosse, S. (2009) Broad-scale determinants of non-native fish species richness are context-dependent. *Proceedings of the Royal Society B: Biological Sciences*, **276**, 2385–2394.

Bomford, M., Kraus, F., Barry, S.C. & Lawrence, E. (2009) Predicting establishment success for alien reptiles and amphibians: a role for climate matching. *Biological Invasions*, **11**, 713–724.

Britton, J.R., Cucherousset, J., Davies, G.D., Godard, M.J. & Copp, G.H. (2010) Non-native fishes and climate change: predicting species responses to warming temperatures in a temperate region. *Freshwater Biology*, **55**, 1130–1141.

Broennimann, O. & Guisan, A. (2008) Predicting current and future biological invasions: both native and invaded ranges matter. *Biology Letters*, **4**, 585–589.

Broennimann, O., Treier, U.A., Müller-Schärer, H., Thuiller, W., Peterson, A.T. & Guisan, A. (2007) Evidence of climatic niche shift during biological invasion. *Ecology Letters*, **10**, 701–709.

Cadotte, M.W., Murray, B.R. & Lovett-Doust, J. (2006) Ecological patterns and biological invasions: using regional species inventories in macroecology. *Biological Invasions*, **8**, 809–821.

Caissie, D. (2006) The thermal regime of rivers: a review. *Freshwater Biology*, **51**, 1389–1406.

Chu, C., Mandrak, N.E. & Minns, C.K. (2005) Potential impacts of climate change on the distributions of several common and rare freshwater fishes in Canada. *Diversity and Distributions*, **11**, 299–310.

Crawford, S.S. & Muir, A.M. (2008) Global introductions of salmon and trout in the genus *Oncorhynchus*: 1870–2007. *Reviews in Fish Biology and Fisheries*, **18**, 313–344.

DAISIE (2009) *Handbook of alien species in Europe*. Springer Netherlands, Dordrecht.

Fausch, K.D., Taniguchi, Y., Nakano, S., Grossman, G.D. & Townsend, C.R. (2001) Flood disturbance regimes influence rainbow trout invasion success among five Holarctic regions. *Ecological Applications*, **11**, 1438–1455.

Fitzpatrick, M.C., Dunn, R.R. & Sanders, N.J. (2008) Data sets matter, but so do evolution and ecology. *Global Ecology and Biogeography*, **17**, 562–565.

Hugueny, B. (1989) West African rivers as biogeographic islands – species richness of fish communities. *Oecologia*, **79**, 236–243.

Hutchinson, G.E. (1957) Concluding remarks. *Cold Spring Harbor Symposia on Quantitative Biology*, **22**, 415–427.

Ineno, T., Tsuchida, S., Kanda, M. & Watabe, S. (2005) Thermal tolerance of a rainbow trout *Oncorhynchus mykiss* strain selected by high-temperature breeding. *Fisheries Science*, **71**, 767–775.

Jackson, D.A. & Harvey, H.H. (1989) Biogeographic associations in fish assemblages: local vs. regional processes. *Ecology*, **70**, 1472–1484.

Jackson, D.A., Peres-Neto, P.R. & Olden, J.D. (2001) What controls who is where in freshwater fish communities – the roles of biotic, abiotic, and spatial factors. *Canadian Journal of Fisheries and Aquatic Sciences*, **58**, 157–170.

Jeschke, J.M. & Strayer, D.L. (2008) Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New York Academy of Sciences*, **1134**, 1–24.

Jiménez-Valverde, A., Lobo, J.M. & Hortal, J. (2008) Not as good as they seem: the importance of concepts in species distribution modelling. *Diversity and Distribution*, **14**, 885–890.

Keith, P. & Allardi, J. (2001) *Atlas des poissons d'eau douce de France*. Muséum National d'Histoire Naturelle, Paris.

Leemans, R. & Cramer, W.P. (1991) *The IIASA database for mean monthly values of temperature, precipitation and cloudiness of a global terrestrial grid*. Report RR-91–18. International Institute for Applied System Analysis (IIASA), Laxenburg.

Leprieur, F., Beauchard, O., Blanchet, S., Oberdorff, T. & Brosse, S. (2008) Fish invasions in the world's river systems: when natural processes are blurred by human activities. *PLoS Biology*, **6**, e28, doi:10.1371/journal.pbio.0060028.

Leprieur, F., Olden, J.D., Lek, S. & Brosse, S. (2009a) Contrasting patterns and mechanisms of spatial turnover for native and exotic freshwater fish in Europe. *Journal of Biogeography*, **36**, 1899–1912.

Leprieur, F., Brosse, S., Garcia-Berthou, E., Oberdorff, T., Olden, J.D. & Townsend, C.R. (2009b) Scientific uncertainty and the assessment of risks posed by non-native freshwater fishes. *Fish and Fisheries*, **10**, 88–97.

Lever, C. (1996) *Naturalized fishes of the world*. Academic Press, London.

Lockwood, J.L., Hoopes, M.F. & Marchetti, M.P. (2007) *Invasion ecology*. Blackwell Scientific, Oxford.

Loo, S.E., Mac Nally, R. & Lake, P.S. (2007) Forecasting New Zealand mudsnail invasion range: model comparisons using native and invaded ranges. *Ecological Applications*, **17**, 181–189.

Matthews, W.J. (1998) *Patterns in freshwater fish ecology*. Chapman and Hall, New York.

Medley, K.A. (2010) Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. *Global Ecology and Biogeography*, **19**, 122–133.

Menke, S.B., Holway, D.A., Fisher, R.N. & Jetz, W. (2009) Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. *Global Ecology and Biogeography*, **18**, 50–63.

Minns, C.K. & Moore, J.E. (1995) Factors limiting the distributions of Ontario's freshwater fish: the role of climate and other variables, and the potential impacts of climate change. *Canadian Journal of Fisheries and Aquatic Sciences*, **121**, 137–160.

Morgan, I.J., D'Cruz, L.M., Dockray, J.J., Linton, T.K., McDonald, D.G. & Wood, C.M. (1998) The effects of elevated winter temperature and sub-lethal pollutants (low pH, elevated ammonia) on protein turnover in the gill and liver of rainbow trout (*Oncorhynchus mykiss*). *Fish Physiology and Biochemistry*, **19**, 377–389.

Moyle, P.B. & Light, T. (1996) Biological invasions of fresh water: empirical rules and assembly theory. *Biological Conservation*, **78**, 149–161.

New, M., Hulme, M. & Jones, P. (1999) Representing twentieth-century space-time climate variability. Part I: development of a 1961–90 mean monthly terrestrial climatology. *Journal of Climate*, **12**, 829–856.

Pearman, P.B., Guisan, A., Broennimann, O. & Randin, C.F. (2008) Niche dynamics in space and time. *Trends in Ecology and Evolution*, **23**, 149–158.

Pyke, G.H. (2008) Plague minnow or mosquito fish? A review of the biology and impacts of introduced *Gambusia* species. *Annual Review of Ecology, Evolution and Systematics*, **39**, 171–191.

Pyšek, P., Richardson, D.M., Pergl, J., Jarošík, V., Sixtová, Z. & Weber, E. (2008) Geographical and taxonomic biases in invasion ecology. *Trends in Ecology and Evolution*, **23**, 237–244.

Rödder, D. & Lötters, S. (2009) Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (*Hemidactylus turcicus*). *Global Ecology and Biogeography*, **18**, 674–687.

Rosenfield, J.A. (2002) Pattern and process in the geographical ranges of freshwater fishes. *Global Ecology and Biogeography*, **11**, 323–332.

Scott, W.B. & Crossman, E.J. (1973) *Freshwater fishes of Canada*. Fisheries Research Board of Canada, Otawa.

Thuiller, W., Richardson, D.M., Pyšek, P., Midgley, G.F., Hughes, G.O. & Rouget, M. (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.

Townsend, C.R. (1996) Invasion biology and ecological impacts of brown trout *Salmo trutta* in New Zealand. *Biological Conservation*, **78**, 13–22.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Figure S1 Climatic range and world-wide distribution of 17 freshwater fish species.

Table S1 Bivariate correlation matrix between the climate variables.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

BIOSKETCHES

Christine Lauzeral is a PhD student in ecology at the University of Toulouse. Her research is focused on predicting the spatial distribution of invasive freshwater fish. She is particularly interested in forecasting species distribution changes under global changes.

Fabien Leprieur is an assistant professor at the Montpellier 2 University. He conducts research on patterns and mechanisms of fish diversity from local to global scales. His research interests include the conservation biogeography of fishes in both freshwater and marine ecosystems.

Editor: Tim Blackburn

Dealing with noisy absences to optimize species distribution models: an iterative ensemble modelling approach

Christine LAUZERAL[1, 2]**, Gaël GRENOUILLET**[1, 2] **& Sébastien BROSSE**[1, 2]

[1] Université de Toulouse, UPS, ENFA; UMR5174 EDB (Laboratoire Évolution et Diversité Biologique); 118 route de Narbonne, F-31062 Toulouse, France.

[2] CNRS; UMR5174 EDB, F-31062 Toulouse, France.

Corresponding author: Christine Lauzeral, Laboratoire Evolution et Diversité Biologique, U.M.R 5174, C.N.R.S - Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex 4, France. Email: christine.lauzeral@univ-tlse3.fr

Summary

Species distribution models (SDMs) are widespread in ecology and conservation biology, but their accuracy can be lowered by non-environmental (noisy) absences that are common in species occurrence data. Here we propose an iterative ensemble modelling (IEM) method to deal with noisy absences and hence improve the predictive reliability of ensemble modelling of species distributions.

In the IEM approach, outputs of a classical ensemble model (EM) were used to update the raw occurrence data. The revised data was then used as input for a new EM run. This process was iterated until the predictions stabilized. The outputs of the iterative method were compared to those of the classical EM using virtual species. The IEM process tended to converge rapidly. It increased the consensus between predictions provided by the different methods as well as between those provided by different learning data sets. Comparing IEM and EM showed that for high levels of non-environmental absences, iterations significantly increased prediction reliability measured by the Kappa and TSS indices, as well as the percentage of well-predicted sites. Compared to EM, IEM also reduced biases in estimates of species prevalence.

Compared to the classical EM method, IEM improves the reliability of species predictions. It particularly deals with noisy absences that are replaced in the data matrices by simulated presences during the iterative modelling process. IEM thus constitutes a promising way to increase the accuracy of EM predictions of difficult-to-detect species, as well as of species that are not in equilibrium with their environment.

Key-words: Bioclimatic models, consensus, ensemble modelling, niche models, species distribution.

Introduction

The ability to predict species distributions is a prerequisite to anticipate environmental changes and to set up sound conservation priorities. There are basically two types of species distribution models (SDMs): mechanistic (or process-based) models that are based on physiological and ecological characteristics of the species and correlative (or niche-based) models that build predictions on the basis of observed species-environment relationships [1]. Mechanistic models require a detailed knowledge of the species considered and are therefore used to predict the distribution of well-known species (e.g., of high conservation or economic value) [reviewed in 2]. In contrast, correlative SDMs are based on the generalization of observed species-environment relationships, and can hence be applied to a large number of species [e.g., 3,4]. For these models, a wide range of predictive statistical methods have been developed since the nineteen eighties. These techniques have been shown to vary considerably in both performance and spatial predictions of species distributions, and despite an abundant literature on method comparisons, no consensus has emerged as to the most suitable statistical method [5-8]. In view of this variability between predictions of SDMs, the recommendation is thus to simultaneously apply a wide range of statistical methods [ensemble modelling, EM, 9].

Presence-absence data are the most commonly used to feed EM as such data are often available over larger areas than species abundances. Although the presence of a species is factual, absence can have a multiple meaning. Lobo *et al.* [10] listed three distinct types of absences: environmental absences (the environmental conditions do not allow the presence of the species), contingent absences (the environmental conditions are favorable but other factors such as biotic interactions, barriers to dispersion or local extinction are responsible for the absence of the species) and methodological absences (the species is present but not detected). Unlike environmental absences (or informative absences), contingent and methodological

absences are noisy absences known to reduce the reliability of SDMs predictions [11]. To account for potential sampling errors and distinguish between non-detection and true absences, binomial likelihood models have been used to estimate changes in range boundaries under recent climate change [12,13] and to correct site-occupancy models for imperfect detection [14]. Although these models are efficient, they are designed to be computed using species abundance data or the detection/non-detection pattern at sites surveyed at least twice [14]. Similarly, Galien et *al.* [15]proposed to combine global- and regional-scale data by weighting pseudo-absences in the regional model to improve SDMs performances. This method is especially efficient for invasive species, that are not in equilibrium with their environment and hence for which contingent absences are frequent. However, this design is only applicable when both large and small scale data are available, which is not the case for most species.

Presence-only SDMs offer another alternative to the problem of absence uncertainty, as they only consider the presence of the species to determine its niche [16,17]. Their performance however remains lower than presence-absence SDMs as they frequently overestimate potential distributions compared with presence–absence models [18]. In order to use presence-absence models when no reliable absence data are available, the use of "pseudo-absences" has also been suggested [18]. "Pseudo-absences" can be simulated through various strategies, but it remains unclear how those strategies affect the models [18-22], so that the use of randomly generated pseudo-absences is often encouraged [22-24]. Such random selection of absences can however reduce model accuracy, leading to an overestimation of the actual range of the species through the selection of uninformative absences, as well as an underestimation of the range through the selection of non-environmental absences (Lobo 2010). Disentangling informative and noisy absences in EM might thus constitute a promising way to enhance the reliability of species distribution predictions.

Here we propose an optimization of EM by using an iterative ensemble model (hereafter called IEM), designed to reduce the effect of noisy absences. To do this, we considered noisy absences to be the false presences predicted by the model (i.e., commission errors, when the model predicted species presence while it was actually absent from the training set). These noisy absences were then considered as presence and the resulting new data matrix was used as a new model training set. This post-processing of model outputs was iterated until the predictions stabilized, therefore providing a potential distribution of the species. Such a strategy presents some similarities with the usual pseudo-absences selection methods [20], but differs by two main points: firstly, it is only based on the use of presence-absence models that are known to be more efficient than presence-only models [18]; secondly, the noisy absences are not discarded but converted into presences.

In this context, the main objectives of this study were: (i) to compare the performances of EM and IEM to predict the spatial distribution of individual species and (ii) to assess the ability of the two modelling methods to deal with noisy absences. To do this, we used simulated occurrence data of three virtual species over France and eight climatic variables. For each species we introduced non-environmental absences in two ways: a random distribution and a distance gradient from the center of the environmental niche. In this last case, the occurrence of non-environmental absences was maximal at the edge of the environmental niche, where the species density usually decreases [25] making the species less detectable.

Material and Methods

Predictor variables

Eight climate variables were extracted over France from the 30” ×30” resolution WorldClim layers for the period 1961-1990 [26]: precipitation in the driest quarter of the year and in the wettest quarter; average monthly precipitation and precipitation seasonality; mean

temperature of the coldest quarter and of the warmest quarter, annual mean temperature and temperature seasonality. These variables were chosen as they are related to the ecological requirements of numerous species, and have often been used in SDMs [27-29].

Virtual ecological niches

The virtual species distributions were defined as hyper volumes of a space defined by a set of relevant environmental variables [30,31]. A normalized principal component analysis (PCA) was computed on the eight climate variables and the first two axes of the PCA, which accounted for 80% of the total variance, were kept as synthetic variables. We hence constructed two synthetic and independent climate variables [31]. In the two-dimensional space created by the two orthogonal axes summarizing climatic variation across France, the virtual species niches were defined as discs [32] centred on (0,0). All geographic cells falling within this disc for the pair of climate variables were considered as the observed distribution range of the virtual species in France. Using three different disc radii, three virtual species were created, with prevalences of 15%, 30% and 60% respectively so as to cover a large prevalence range.

Data sets

First, 1000 cells were randomly sampled among the 912730 cells covering the entire surface of France. These 1000 cells were considered as the sampling sites. This operation was repeated 100 times, giving rise to 100 data sets. Each of these 100 data sets was randomly split into two parts: two-thirds of the data were used to calibrate the SDMs and the remaining third was used as a test set.

Then, five occurrence levels of noisy absences (15%; 30%; 45%; 60% and 75% of all the presences available in the learning data set) were inserted into the learning data set. For each occurrence level, two strategies were used to determine the position of the noisy absences. On the one hand, noisy absences were selected randomly from all the presences available in the

learning data set. On the other hand, we assumed that the probability of a site inside the niche to be a noisy absence increased as a Gaussian function of the distance to the centre of the environmental niche. More explicitly, the probability of the site being selected was equal to $(1-0.9\ \exp^{(-d^2/r^2)})/n$ where d was the distance to the centre of the environmental niche, r was the radius of the environmental niche and n was chosen to ensure that the sum of the probabilities over all presence sites is equal to 1. We thus obtained 1000 (5 noisy absence percentages, 2 absence distribution types, 100 repetitions) data sets for each species.

IEM modelling

According to the EM framework, we used six predictive modelling methods: generalized linear models (GLM); generalized additive models (GAM); boosted trees (BT); classification and regression trees (CART); generalized boosted regression models (GBM) and linear discriminant analysis (LDA). For the GLM and LDA models, squared variables were included in the model to deal with non-linearity. The modelling followed the classical EM process. At the first iteration, the learning data was the original data with n = 666 sites (Step 1; Fig. 1). The six statistical methods were used to build models using this learning data set (Step 2; Fig. 1). For each site, the six resulting probabilities of presence (one per modelling method) were then averaged, giving rise to a per-site probability of presence [33] (Step 3; Fig. 1). We refrained from weighting the six model outputs using an accuracy measurement like the AUC, because the data set contained noisy absences. Indeed, weighting the outputs of the modelling methods could favour the models that overfit the data and hence reduce the correction rate of noisy absences. Lastly, the probability vector was converted into a presence-absence response, using a cut-off threshold maximizing the Kappa index (Step 4; Fig. 1). This approach was preferred to the ROC curve approach (maximising the sum of sensitivity and specificity) that gives less accurate prevalence predictions [34,35]. These four steps corresponded to a classical EM procedure, and account for one IEM iteration. The predicted data matrix

obtained at the end of the first iteration was used to update the raw data set before the next iteration. To do this, observed and predicted data matrices were compared and an absence was considered as noisy when the model predicted presence while the species was absent from the observed data (i.e. commission error). In that case, we updated the raw data by replacing absence (0) by presence (1). The resulting data matrix was then used as the learning data set for the following iteration (Step 1; Fig. 1). The entire procedure was then repeated 100 times (Fig. 1). The modelling procedure was implemented in R [36].

Models variability

To evaluate the prediction variability inherent to the statistical methods (i.e., GBM and BT), we ran the EM 100 times for each species and each complete data set. We observed that in 95% of the cases less than 5% of the 334 test sites had variable predictions (and 11% of the sites had variable predictions). The number of different predictions was less than 27 in 95% of the cases. We thus considered that our IEM model had stabilized when less than 5% of the sites provided variable predictions in 27 successive iterations.

The evolution of the variability among the six SDM predictions through the iterative process was evaluated at each iteration. Following Thuiller [37], we carried out a standardized Principal Component Analysis (PCA) on the data matrix made up of the 6 probability-of-presence vectors at the 334 test sites, and we evaluated the consensus among the predictions by calculating the percentage of variance accounted for by the first axis of the PCA.

The variability of the EM and IEM binary predictions inherent to the sampling of learning sites was evaluated in the same way for each virtual species, each percentage of noisy absences and each absence selection. As the 100 tests sets share a very low number of cells, we randomly selected 1000 cells among the 912730 cells covering the entire surface of France. These cells were used as a common test set for the 100 models built on the 100 learning data sets. For each of the 100 models, we predicted the presence-absence of the

species over these 1000 cells. Then, we carried out a PCA on the data matrix made up of the 100 presence-absence vectors. The consensus among the predictions was evaluated by calculating the percentage of variance accounted for by the first axis of the PCA.

Comparing IEM and EM

For each of the species, we first evaluated the AUC [38] of the mean model on the 334 test sites before using the Kappa cut-off threshold. We then evaluated the predictive accuracy of both EM and IEM presence-absence predictions on the test sites by measuring three complementary and commonly used indicators: (i) the percentage of correctly predicted sites that provides a direct measure of both true absences and true presences; (ii) the Kappa index; its dependence on prevalence merely reflects its role as a chance-corrected measure [39]; and (iii) the True Skill Statistic (TSS) which is more independent of observed species prevalence than Kappa [40]. As a complement, we assessed the ability of EM and IEM to predict the prevalence of the species by measuring the difference between the observed and the predicted prevalences. Pairwise comparisons between EM and IEM were done using Wilcoxon's tests.

Finally, we used a null-model simulation to explore the possibility that the increase in model accuracy between EM and IEM could only be due to an increase in the predicted prevalence through the iterative process. We hence compared IEM predictive accuracy to the accuracy of the output of EM predictions modified by randomly turning some sites from absences to presences. The number of sites where absences were replaced by presences was identical to that turned from absences to presences by the IEM procedure, and we computed the percentage of mispredicted sites, Kappa and TSS. For each species, we reiterated this procedure 10 000 times for each learning data set and we compared the observed values of indices produced by the IEM with the distribution of the 10 000 values simulated by the null-model.

We also plotted a map of omission and commission errors. For each species, we predicted the presence-absence of the species over the 912730 cells covering the French territory. We then counted, over the 100 models built using the 100 different learning data sets, the percentage of mispredicting models in each cell. This was done for both EM and IEM.

Results

IEM modelling

For the three species, the iterative process tended to converge rapidly, as most of the predictions stabilized after 2 to 70 iterations (mean: 15 iterations). Only 3.5% of the models did not stabilize after 70 iterations (see Fig. S1 a). The models that did not stabilize were characterized by high levels of noisy absences. Moreover, the stabilization time (i.e. number of iterations) increased with the percentage of noisy absences (Fig. S1 b).

After a few iterations, the 6 different methods provided consensual predictions for the 334 test sites (Fig. S2). At the first iteration (i.e., EM), the mean percentages of variance accounted for by the first axis of the PCA were 67.5%, 72.4% and 79.5% for the three species, respectively. Using IEM, consensus increased after 25 iterations up to 73.9%, 79%, and 86.1% respectively and then reached a relatively stable plateau up to the end of the iterative procedure (Fig. S2).

IEM also increased the consensus of predictions built on different learning data sets (Fig. S3). At the first iteration (i.e., the EM), the mean percentages of variance accounted for by the first axis of the PCA were 81.5%, 72.2%, 59.6%, 45.9% and 32.3% respectively for the five noisy absence levels. Using IEM, consensus increased up to 82%, 79.5%, 74.5%, 66.7% and 49.6% respectively. This increase was higher for frequent species especially when noisy absences were randomly selected.

Predictive performance

The AUC almost always significantly decreased during the iterative process but this decrease remained low except for high levels of noisy absences (Fig. 2). All AUC values were higher than 0.77 (higher than 0.88 for 95% of the models) for IEM whereas they were higher than 0.79 (higher than 0.92 for 95% of the models) for EM. Evaluating the predictive accuracy of both EM and IEM presence-absence output on the 334 test sites showed that compared with EM, IEM significantly reduced false absences (Wilcoxon test, $p<0.001$, Fig. 3). Due to the IEM principle (i.e., replacing noisy absences by presences in the learning data set), the model most easily predicted presences in environments that were in fact true absences, and hence false presences increased significantly in the test set predictions (Wilcoxon test, $p<0.001$, Fig. 3). Lowering false absences and increasing false presences led to a variation of the predictive accuracy evaluated on the test set that almost depended on the percentage of noisy absences (Fig. 3). Using IEM, the three species experienced a significant increase in predictive accuracy for noisy absences levels greater than 30% (Wilcoxon test, $p<0.001$). The results were more mixed for lower levels of noisy absences (15 an 30%) as both positive, negative or no change were detected between EM and IEM according to the quality index. Although some were significant, these changes remained of slight intensity (Fig. 3).

For noisy absence levels greater than 30%, iterations increased the percentage of well-predicted sites, Kappa and TSS in 93%, 97% and 84% of the cases, respectively (Fig. 3). Moreover, the Kappa index calculated for IEM gave a good score (>0.6) for 2253 out of the 3000 cases and a moderate score (between 0.4 and 0.6) for 593 cases. Our predictions were thus reliable (i.e. Kappa > 0.4) in 94.9% of the cases. The performance of EM was clearly lower, with only 1376 cases reaching a Kappa score above 0.6, and 71% of the cases for which the predictions were reliable. The TSS index confirmed this trend, as TSS calculated for IEM reached a score greater than 0.6 for 73.7% of the cases and between 0.4 and 0.6 for

18.6% of the cases. TSS was lower for EM with a score greater than 0.6 for 38.8% of the cases and a score between 0.4 and 0.6 for 27% of the cases. Moreover, the IEM provided less biased estimates of species prevalence in 80.7% of the cases (96.5% of the cases with noisy absences levels greater than 30%) (Fig. 3). Note that for high levels of noisy absences, the benefit of IEM compared to EM was lower if the noisy absences were preferentially located at the edge of the niche. Moreover, species prevalence only affected the pattern at the highest noisy absence level due to the limited increase in model quality through iterations for the rarest species.

The geographical pattern of omission errors depended on the selection of noisy absences. When the noisy absences were randomly chosen, the EM mispredicted presence cells spread over the whole distribution and were slightly more abundant at the edge of the distribution. For abundant, non-random noisy absences, the EM only predicted the 'core' region of each species distribution. At the end of the iterative process, the remaining often omitted cells were in both cases more abundant at the edge of the distribution, but this pattern was less marked for randomly chosen noisy absences (Fig. 4, S4, S5).

The location of commission errors was less affected by the selection of noisy absences. Mispredicted absence sites were mostly located at the edge of the distribution and the IEM increased the mispredicted areas especially in areas where the environmental variables varied only slightly (Fig. 4, S4, S5).

The increase in the predictive performance between IEM and EM was not due to the rise of the predicted prevalence, as for 2963 out of the 3000 cases (1789 of the 1800 cases with levels of noisy absences greater than 30%), IEM predictions were significantly more reliable than those produced by the null-model simulations ($p<0.05$), considering TSS, Kappa and the percentage of well-predicted sites (Table 1).

Discussion

In ecological sciences, the presence of an organism is factual while absence is inferred i.e., the species was not seen or identified or captured [10]. Absence is hence the main cause of uncertainty in species occurrence data matrices and thus can have detrimental consequences on the relevance of SDMs [10,11]. Alternatives are limited as the only models currently used to take species detectability into account require repeated survey data or species abundance data [14], while most data matrices are composed of presence-absence data without multiple observations. IEM provides a way to reduce this problem as it only requires presence-absence data matrices and can reduce the bias inherent in species detectability by dealing with noisy absences. Although IEM did not provide better results than the EM with low levels of noisy absences, it was significantly more efficient than EM as soon as the data set contained more than 30% of noisy absences. In such cases, it enhanced the prediction ability of SDMs by increasing both the quality of the statistical models and the consensus between statistical methods. This is an important point as the variability between statistical models is recognized as the major source of uncertainty in the prediction of species spatial distributions by SDMs [27]. This tendency is triggered for low detectable species [41,42], such as species of low occurrence like large predators in forested areas [e.g. 43]. In the same way, threatened species are characterized by a high occurrence of non-contingent absences as those species have often been extirpated from a large part of their natural area. The IEM approach might therefore be of interest in the prediction of the potential distribution of threatened or difficult-to-detect species, which is not readily feasible using classical SDMs [44].

Another possible application of the IEM is the prediction of the potential distribution of non-native species, which has been considered as difficult to achieve using SDMs [e.g., 45,46]. Indeed, it is now recognised that most non-native species are in a non-equilibrium state, particularly due to spatial variability in propagule pressure and human impact on ecosystems

across the world [47,48]. Up until now the two ways proposed to predict the spatial invasion range of invasive species involved (1) the use of presence-only models, which have a low predictive efficiency [49,50], or (2) the calibration of models on the niche conditions found in both the native and the exotic range of the species [21,51,e.g., 52], with the aim of accounting for potential niche shifts between native and invasion ranges [52-55]. This however strongly limits the predictive efficiency of the models, as a substantial part of the absences in the exotic range are contingent, leading to an underprediction of the potential range. The IEM might hence constitute an alternative for predicting the invasion potential of current and future invaders as it has been shown to reduce omission errors that are known to be costly in the prediction of invasive species distribution, as it is more difficult to eradicate a pest than to identify a species that may become a problem [56].

As for more classical SDMs, the spatial extent of presence data remains determinant in the quality of species predictions. Although IEM has been shown to be of interest in the reduction of omission errors, it should however be noted that this method remains unable to guess missing ecological information. This was observed in two ways on our virtual species for high levels of false absences. First, IEM showed less improvement in the accuracy of models built on data sets with noisy absences located at the edge of the niche. As the model did not have information on suitable environmental conditions at the edge of the environmental niche, it tended to underpredict the distribution range. Second, rare species experienced a lower accuracy increase through iteration for the highest level of noisy absence. Here, the number of observed occurrences probably fell under a critical threshold that did not permit the models to gather sufficient information to build a detailed image of the niche.

Particular attention should also be given to the selection of the environmental variables, which always remains a crucial point in the model building process [24,51]. This is particularly true for IEM as an inaccurate variable may drive predictions in the wrong direction through

iterations. For the same reason, we also warn against the use of the iterative approach when using a unique statistical method as iterations may increase bias inherent to the statistical method used, whereas the use of ensemble methods buffers potential bias due to any specific statistical method [57].

As IEM is designed to fill noisy absences, it may also be affected by false presences (the species has been detected outside its niche). IEM might then inflate the niche by considering as noisy absences those falling in the gap between real and false presences. False presences might therefore promote IEM niche overprediction or drive the model in the wrong direction, especially in the case of high levels of noisy absences that give more importance to the false presences. Although false presences are usually rare in ecological data, they can occur as species misidentification in data bases or as recorded occurrences of non-established species (i.e. the species is recorded in an environmental niche where it is unable to settle). The effect of false presences on IEM hence deserves to be quantified.

In the same way, model transferability should be evaluated. We showed here that compared to EM, IEM increased the consensus between predictions based on different learning data sets. This suggests that IEM tends to reduce both the sensitivity of models to differences in the ranges of environmental predictors and the overfitting of the learning data. As these two parameters are known to reduce model transferability [58], IEM might be more transferable than EM. But EM and IEM transferability remains to be compared on real species as numerous ecological parameters are known to affect model transferability [58].

Finally, many parameters are known to affect the quality of SDMs, such as the size and extent of the observed distribution, environmental parameter sampling [59], the prevalence of the species [60], cut-off selection [61], or the selection of absences used in the learning data set. The sensitivity of the IEM to these parameters remains to be evaluated before intensively

using IEM. We therefore encourage complementary studies to draw up precise guidelines for the use of this method.

Acknowledgements

EDB is part of the "Laboratoire d'Excellence" (LABEX) entitled TULIP (ANR-10-LABX-41).

References

1. Morin X, Thuiller W (2009) Comparing niche- and process-based models to reduce prediction uncertainty in species range shifts under climate change. Ecology 90: 1301-1313.
2. Kearney M, Porter W (2009) Mechanistic niche modelling: combining physiological and spatial data to predict species' ranges. Ecol Lett 12: 334-350.
3. Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8: 993-1009.
4. Elith J, Graham CH, Anderson RP, Dudik M, Ferrier S, et al. (2006) Novel methods improve prediction of species' distributions from occurrence data. Ecography 29: 129-151.
5. Olden JD, Jackson DA (2002) A comparison of statistical approaches for modelling fish species distributions. Freshwater Biol 47: 1976-1995.
6. Segurado P, Araújo MB (2004) An evaluation of methods for modelling species distributions. J Biogeogr 31: 1555-1568.
7. Manel S, Dias JM, Ormerod SJ (1999) Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird. Ecol Model 120: 337-347.
8. Elith J, Leathwick JR (2009) Species distribution models: Ecological explanation and prediction across space and time. Annu Rev Ecol Evol S 40: 677-697.
9. Araújo MB, Whittaker RJ, Ladle RJ, Erhard M (2005) Reducing uncertainty in projections of extinction risk from climate change. Global Ecol Biogeogr 14: 529–538.
10. Lobo JM, Jiménez-Valverde A, Hortal J (2010) The uncertain nature of absences and their importance in species distribution modelling. Ecography 33: 103-114.
11. Lobo JM (2008) More complex distribution models or more representative data? Biodiversity Informatics 5: 14-19.
12. Rowe RJ, Finarelli JA, Rickart EA (2010) Range dynamics of small mammals along an elevational gradient over an 80-year interval. Glob Change Biol 16: 2930-2943.
13. Moritz C, Patton JL, Conroy CJ, Parra JL, White GC, et al. (2008) Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. Science 322: 261-264.
14. Kéry M, Gardner B, Monnerat C (2010) Predicting species distributions from checklist data using site-occupancy models. J Biogeogr 37: 1851-1862.
15. Gallien L, Douzet R, Pratte S, Zimmermann NE, Thuiller W (2012) Invasive species distribution models - how violating the equilibrium assumption can create new insights. Global Ecol Biogeogr, *in press*.
16. Hirzel AH, Hausser J, Chessel D, Perrin N (2002) Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? Ecology 83: 2027-2036.
17. Farber O, Kadmon R (2003) Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. Ecol Model 160: 115-130.
18. Zaniewski AE, Lehmann A, Overton JM (2002) Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol Model 157: 261-280.
19. Chefaoui RM, Lobo JM (2008) Assessing the effects of pseudo-absences on predictive distribution model performance. Ecol Model 210: 478-486.

20. Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J Appl Ecol 41: 263-274.
21. Capinha C, Leung B, Anastácio P (2011) Predicting worldwide invasiveness for four major problematic decapods: an evaluation of using different calibration sets. Ecography 34: 448-459.
22. Stokland JN, Halvorsen R, Støa B (2011) Species distribution modelling-Effect of design and sample size of pseudo-absence observations. Ecol Model 222: 1800–1809.
23. Kadmon R, Farber O, Danin A (2003) A systematic analysis of factors affecting the performance of climatic envelope models. Ecol Appl 13: 853-867.
24. Wisz MS, Guisan A (2009) Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. BMC Ecol 9: 8.
25. Brown JH (1984) On the relationship between abundance and distribution species. Am Nat 124: 225-279.
26. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. Int J Climatol 25: 1965-1978.
27. Buisson L, Thuiller W, Casajus N, Lek S, Grenouillet G (2010) Uncertainty in ensemble forecasting of species distribution. Glob Change Biol 16: 1145-1157.
28. Thuiller W, Lavorel S, Araújo MB, Sykes MT, Prentice IC (2005) Climate change threats to plant diversity in Europe. P Natl A Sci USA 102: 8245-8250.
29. Marini MA, Barbet-Massin M, Lopes LE, Jiguet F (2009) Predicted climate-driven bird distribution changes and forecasted conservation conflicts in a Neotropical savanna. Conserv Biol 23: 1558-1567.
30. Lobo JM, Tognelli MF (2011) Exploring the effects of quantity and location of pseudo-absences and sampling biases on the performance of distribution models with limited point occurrence data. J Nat Conserv 19: 1-7.
31. Jiménez-Valverde A, Lobo JM (2007) Threshold criteria for conversion of probability of species presence to either-or presence-absence. Acta Oecol 31: 361-369.
32. Soberón J, Nakamura M (2009) Niches and distributional areas: Concepts, methods, and assumptions. P Natl A Sci USA 106: 19644–19650.
33. Marmion M, Parviainen M, Luoto M, Heikkinen RK, Thuiller W (2009) Evaluation of consensus methods in predictive species distribution modelling. Divers Distrib 15: 59-69.
34. Freeman EA, Moisen GG (2008) A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol Model 217: 48-58.
35. Mouton AM, De Baets B, Van Broekhoven E, Goethals PLM (2009) Prevalence-adjusted optimisation of fuzzy models for species distribution. Ecol Model 220: 1776-1786.
36. R Development Core Team (2011) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
37. Thuiller W (2004) Patterns and uncertainties of species' range shifts under climate change. Glob Change Biol 10: 2020–2027.
38. Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24: 38-49.
39. Santika T (2011) Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. Global Ecol Biogeogr 20: 181-192.

40. Allouche O, Tsoar A, Kadmo R (2006) Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J Appl Ecol 43: 1223-1232.
41. Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, et al. (2008) Effects of sample size on the performance of species distribution models. Divers and Distrib 14: 763-773.
42. Pearson RG, Raxworthy CJ, Nakamura M, Peterson AT (2007) Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. J Biogeogr 34: 102-117.
43. Cubaynes S, Pradel R, Choquet R, Duchamp C, Gaillard J-M, et al. (2010) Importance of accounting for detection heterogeneity when estimating abundance: the case of french wolves. Conserv Biol 24: 621-626.
44. Cianfrani C, Le Lay G, Hirzel AH, Loy A (2010) Do habitat suitability models reliably predict the recovery areas of threatened species? J Appl Ecol 47: 421-430.
45. Gallien L, Munkemuller T, Albert CH, Boulangeat I, Thuiller W (2010) Predicting potential distributions of invasive species: where to go from here? Divers Distrib 16: 331-342.
46. Olden JD, Kennard MJ, Leprieur F, Tedesco PAW, K.O., Garcia-Berthou E (2010) Conservation biogeography of freshwater fishes: recent progress and future challenges. Divers Distrib 16: 496-513.
47. Leprieur F, Beauchard O, Blanchet S, Oberdorff T, Brosse S (2008) Fish invasions in the world's river systems: when natural processes are blurred by human activities PloS Biol 6: e28. http://dx.doi.org/10.1371/journal.pbio.0060028.
48. Blanchet S, Leprieur F, Beauchard O, Staes J, Oberdorff T, et al. (2009) Broad-scale determinants of non-native fish species richness are context-dependent. P Roy Soc B-Biol Sci 276: 2385-2394.
49. Václavík T, Meentemeyer RK (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? Ecol Model 220: 3248-3258.
50. Brotons L, Thuiller W, Araújo MB, Hirzel AH (2004) Presence-absence versus presence-only modelling methods for predicting bird habitat suitability. Ecography 27: 437-448.
51. Jiménez-Valverde A, Peterson AT, J. Soberón J, Overton JM, Aragón P, et al. (2011) Use of niche models in invasive species risk assessments. Biol Invasions 13: 2785-2797.
52. Beaumont LJ, Gallagher RV, Thuiller W, Downey PO, Leishman MR, et al. (2009) Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. Divers Distrib 15: 409-420.
53. Lauzeral C, Leprieur F, Beauchard O, Duron Q, Oberdorff T, et al. (2011) Identifying climatic niche shifts using coarse-grained occurrence data: a test with non-native freshwater fish. Global Ecol Biogeogr 20: 407-414.
54. Rödder D, Lötters S (2009) Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (Hemidactylus turcicus). Global Ecol Biogeogr 18: 674-687.
55. Medley KA (2010) Niche shifts during the global invasion of the Asian tiger mosquito, Aedes albopictus Skuse (Culicidae), revealed by reciprocal distribution models. Global Ecol Biogeogr 19: 122-133.
56. Mack RN, Simberloff D, Lonsdale WM, Evans H, Clout M, et al. (2000) Biotic invasions: Causes, epidemiology, global consequences, and control. Ecol Appl 10: 689-710.
57. Araújo MB, New M (2007) Ensemble forecasting of species distributions. Trends Ecol Evol 22: 42-47.

58. Randin CF, Dirnbo T, Dullinger S, Zimmermann NE, Zappa M, et al. (2006) Are niche-based species distribution models transferable in space? J Biogeogr 33: 1689–1703.
59. Menke SB, Holway DA, Fisher RN, Jetz W (2009) Characterizing and predicting species distributions across environments and scales: Argentine ant occurrences in the eye of the beholder. Global Ecol Biogeogr 18: 50-63.
60. Williams JN, Seo C, Thorne J, Nelson JK, Erwin S, et al. (2009) Using species distribution models to predict new occurrences for rare plants. Divers Distrib 15: 565-576.
61. Liu CR, Berry PM, Dawson TP, Pearson RG (2005) Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28: 385-393.

Figures:

Figure 1: The iterative ensemble modelling (IEM) process. Step 1: At the first iteration, the learning data is the original data set with n = 666 sites. For the following iterations, the learning data is the raw data set updated using the predicted data matrix: an absence is considered as noisy if the model predicts presence while the species is absent from the observed data. In that case, the raw data is updated by replacing absence (0) by presence (1); Step 2: The six statistical methods are used to build models with the learning data set; Step 3: the six resulting probabilities of presence for each site (one per modelling method) are averaged, giving rise to a per-site probability of presence; Step 4: the probability vector is converted into a presence-absence response, using a cut-off threshold maximizing the Kappa index.

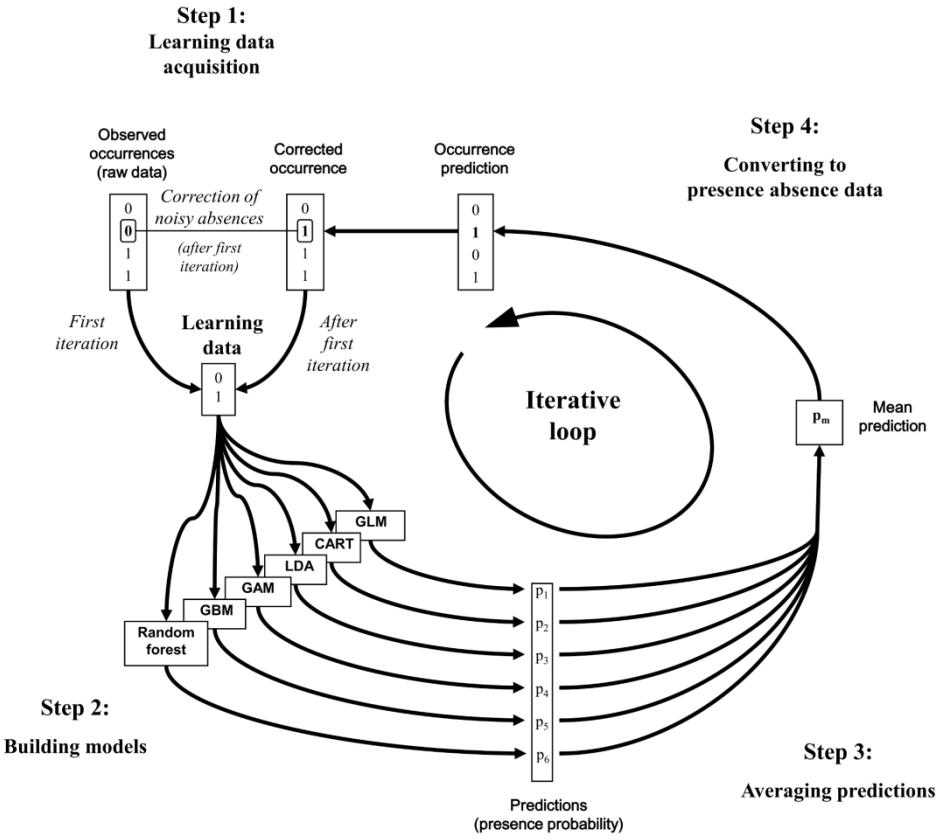


Figure 2: Effects of noisy absences on threshold-independent measurements (i.e., AUC) of model accuracy after the first iteration (EM, in green) and at the end of the process (IEM, in black) for three virtual species with true prevalence of (A) 15%, (B) 30% and (C) 60%. Box colors represent geographic distribution of noisy absences (grey: random; white: mostly at the edge of the niche).

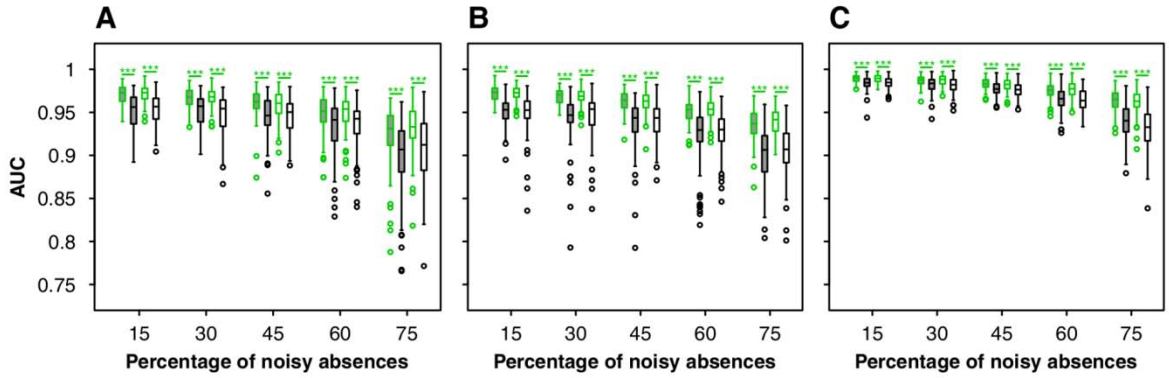


Figure 3: Effects of noisy absences on threshold-dependent measures of model accuracy after the first iteration (EM, in green) and at the end of the process (IEM, in black) for three virtual species with true prevalence of (A) 15%, (B) 30% and (C) 60%. Model accuracy was evaluated using the two types of mispredicted sites, percentage of well-predicted sites, TSS, Kappa, and predicted prevalence. Box colors represent geographic distribution of noisy absences (grey: random; white: mostly at the edge of the niche). The grey line corresponds to the true value of the prevalence.

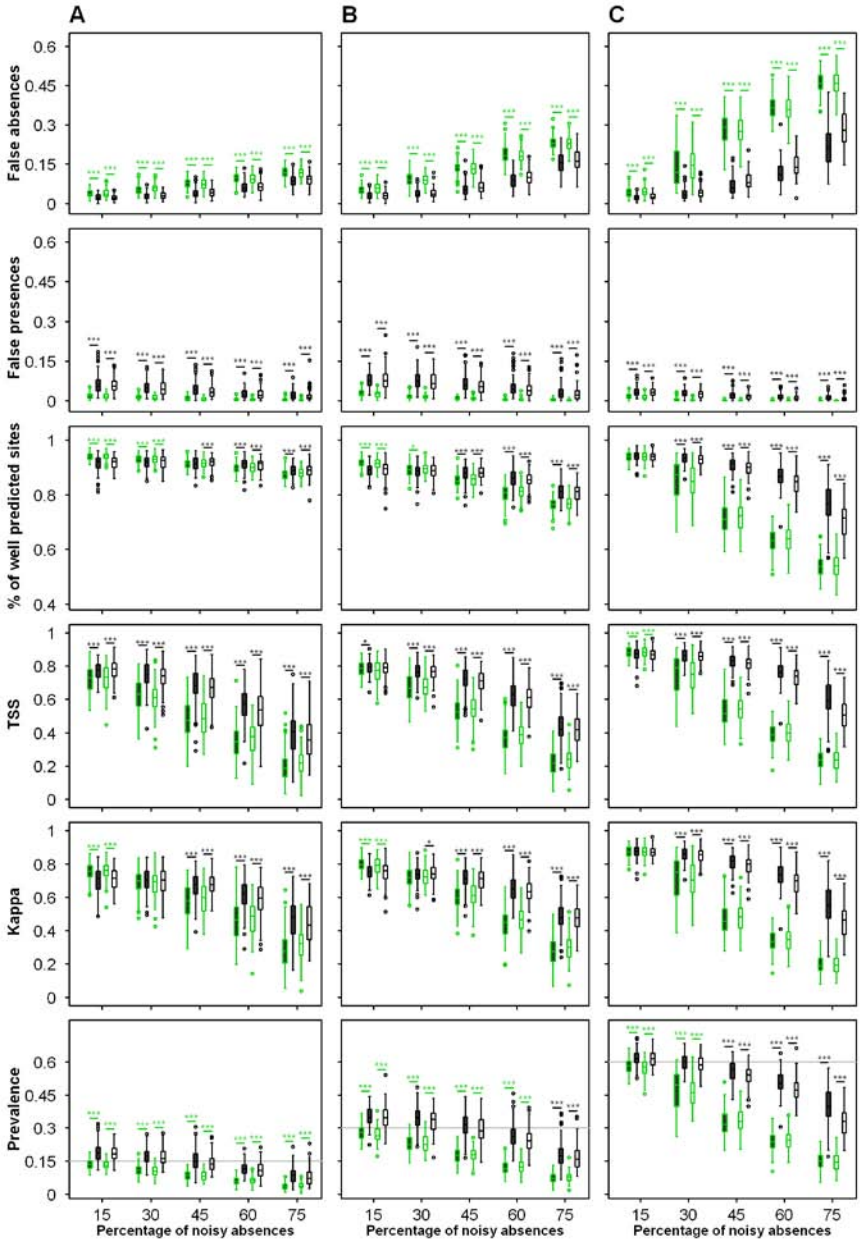


Figure 4: (A) The observed and (B) predicted distributions of the frequent species (prevalence = 60%) using noisy absences randomly located or (C) located following a distance gradient from the center of the environmental niche. For each noisy absence type, the top line of 3 maps refers to EM and the bottom line maps to IEM. For each line, noisy absences increase from the left to the right (left 15%, centre 45%, right 75%). The situations with 30 % and 60% of noisy absences are not shown for clarity. The 100 models based on the 100 different learning data sets were used and we evaluated the percentage of mispredicting models in each pixel. The darker the pixels, the higher the percentage of prediction errors.

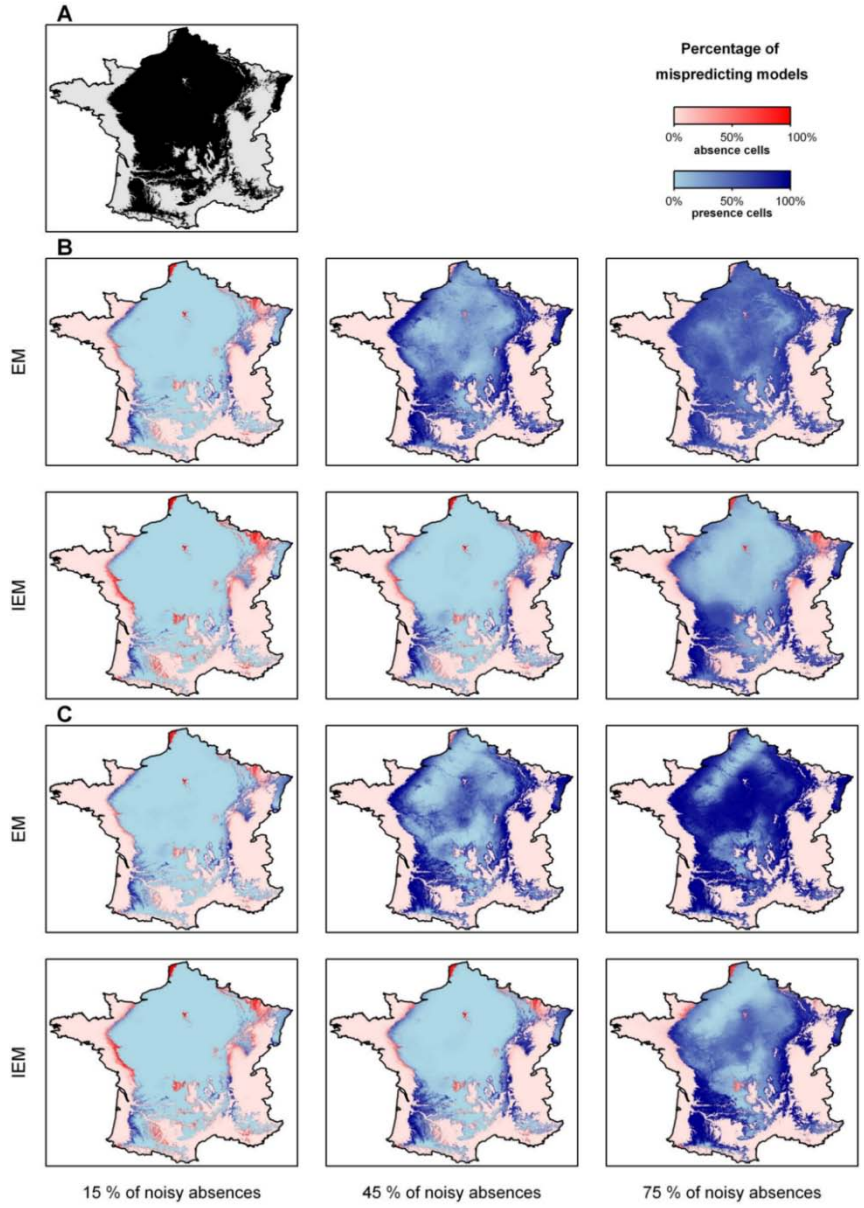


Table 1: Null-model simulations. Number of IEM models with accuracy not better than expected by chance among the 100 built on the 100 learning data sets. We counted the number of sites turned from absences to presences by the IEM procedure. The same number of sites predicted as absences by EM were randomly selected and replaced by presences. The accuracy of the resulting model was evaluated using the percentage of well-predicted sites, the TSS and Kappa indices. The accuracy was considered as lower if at least one of the three indices of the IEM was lower than that evaluated on the random predictions. The random sampling was repeated 10000 times.

		Random location of noisy absences			**Gradient of noisy absences from the center of the niche**		
	Percentage of noisy absences	$0.01 \leq p < 0.05$	$0.01 \leq p < 0.01$	$p < 0.001$	$0.01 \leq p < 0.05$	$0.01 \leq p < 0.01$	$p < 0.001$
Prevalence 15%	15%	7	8	9	5	5	15
	30%	3	0	4	1	1	8
	45%	2	0	3	1	1	4
	60%	1	2	4	2	3	3
	75%	3	1	7	2	2	11
Prevalence 30%	15%	2	5	7	1	4	5
	30%	0	1	3	2	0	1
	45%	0	0	2	0	0	1
	60%	0	0	0	0	0	0
	75%	0	1	1	0	0	1
Prevalence 60%	15%	2	1	9	3	1	11
	30%	0	0	1	0	0	0
	45%	0	0	0	0	0	0
	60%	0	0	0	0	0	0
	75%	0	0	0	0	0	0

Supplementary material:

Figure S1: Stabilization of the iterative process. a) Number of sites with variable predictions during the 27 following iterations. The grey line corresponds to maximum value over the 3000 models, vertical bars correspond to the variability across 95% of the models; dots correspond to the mean values across the 3000 models. The two dashed lines correspond to the variability inherent to the statistical methods (for all the simulations and for the 95% less variable ones). b) Stabilization time (in number of iterations) of the iterative process across noisy absence levels.

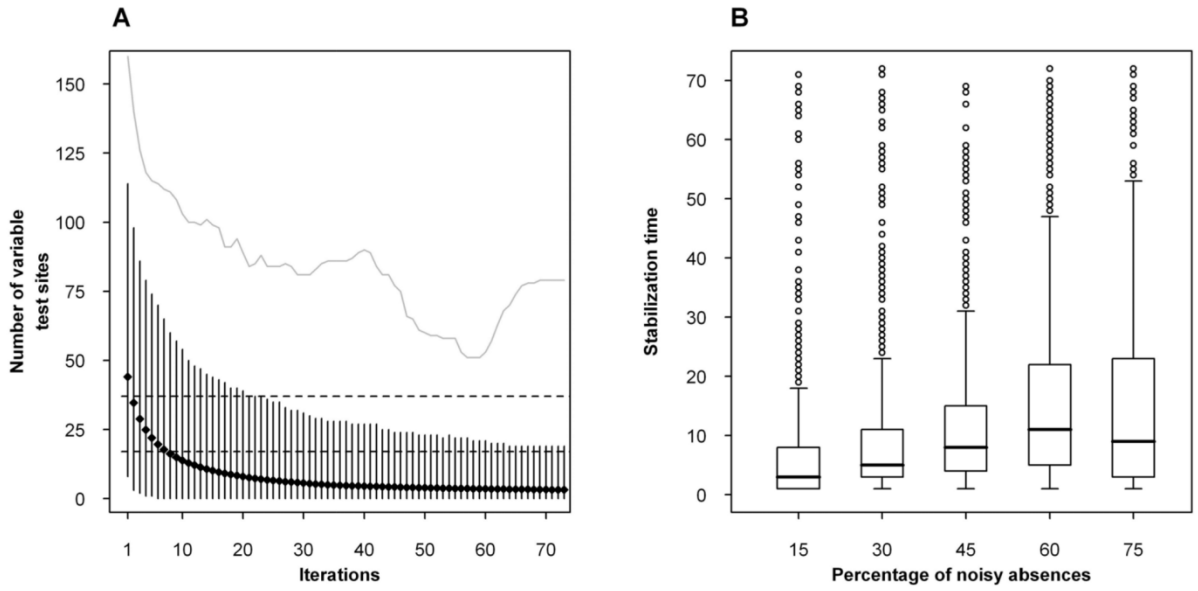


Figure S2: Consensus (percentage of variance explained by the first axis of the PCA) among the six models during the iterative process for the 334 test sites. Species prevalence (a) 15%; (b) 30%; (c) 60%. Grey lines correspond to maximum and minimum values, vertical bars correspond to the variability across 95% of the test sites; dots correspond to the mean variance.

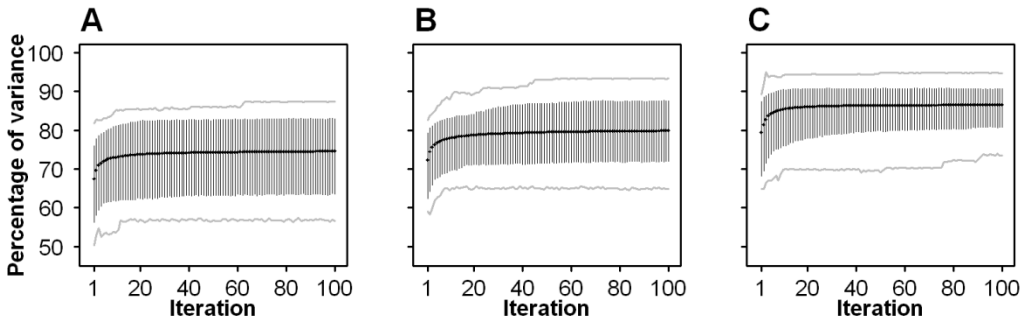


Figure S3: Consensus (percentage of variance explained by the first axis of the PCA) among the 100 learning data sets after the first iteration (EM) and at the end of the process (IEM) for 1000 randomly selected cells over France. Symbols represent virtual species prevalence. Circles: 15%; squares: 30%; diamonds: 60%. Colour represents noisy absence samplings. Grey: random; white: almost at the edge of the niche. Border colour represents the models. Green: EM; black: IEM.

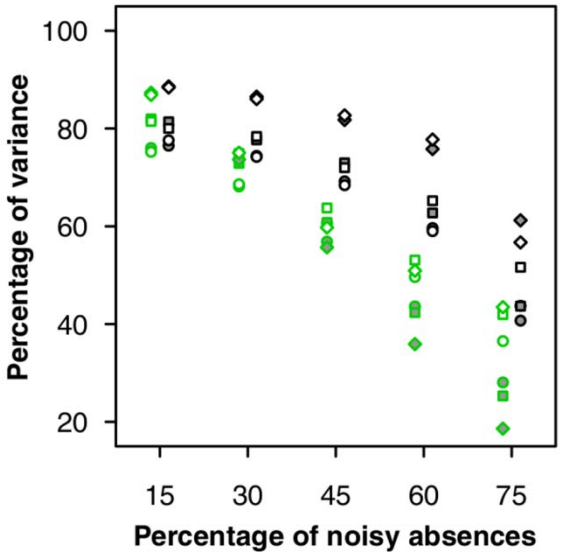


Figure S4: The (a) observed and (b) predicted distributions of the rare species (prevalence = 15%) using noisy absences randomly located or (c) located following a distance gradient from the center of the environmental niche. For each noisy absence type, the top line of 3 maps refers to EM and the bottom line maps to IEM. For each line, noisy absences increase from left to right (left 15%, centre 45%, right 75%). The situations with 30 % and 60% of noisy absences are not shown for clarity. The 100 models based on the 100 different learning data sets were used and we evaluated the percentage of mispredicting models in each pixel. The darker the pixels, the higher the percentage of prediction errors.

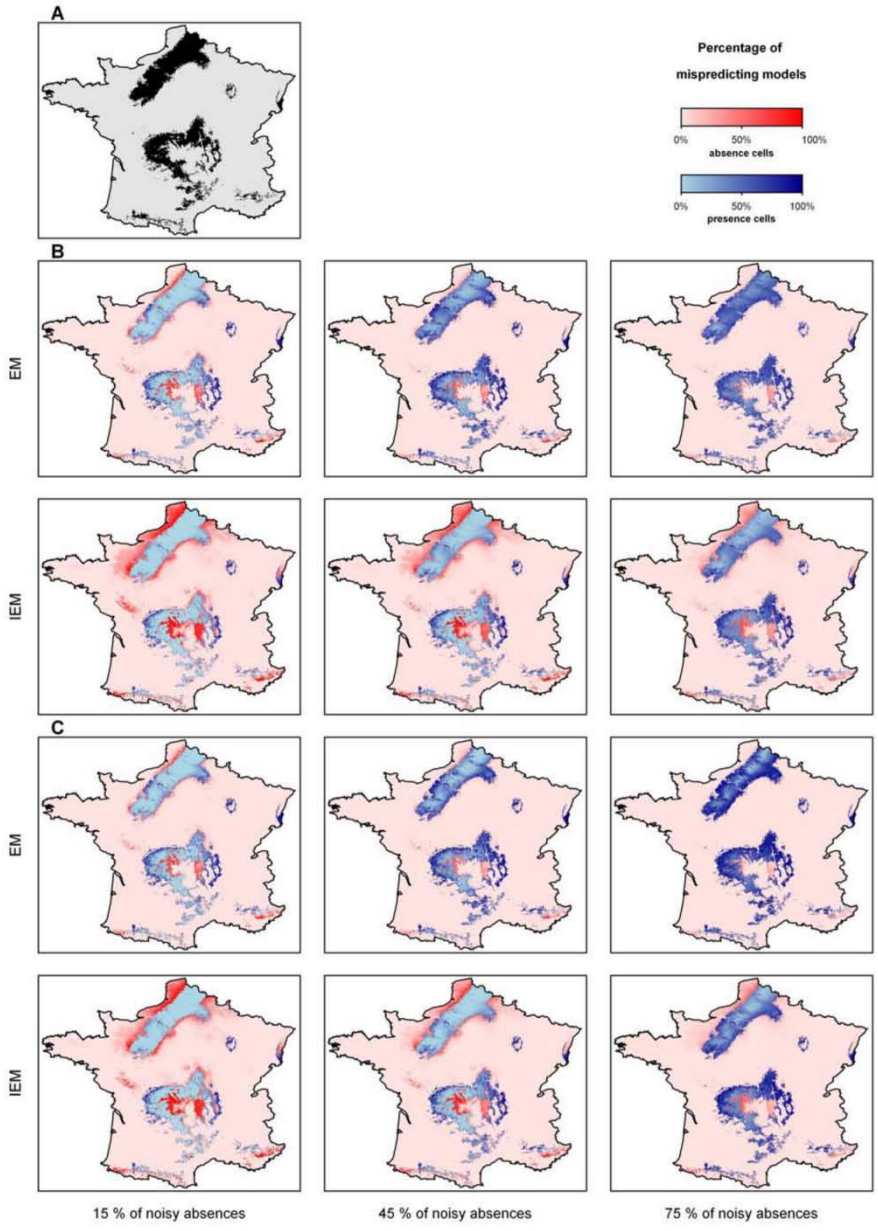
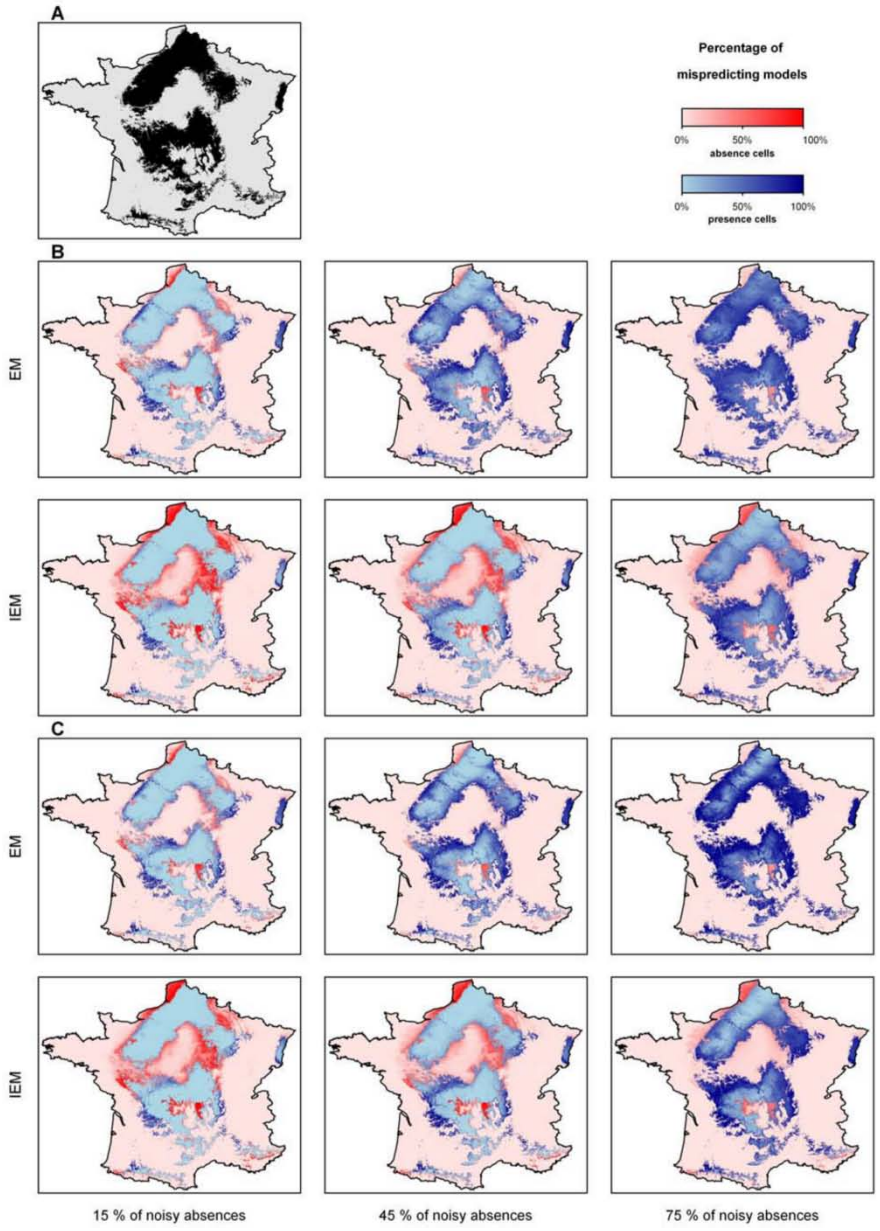


Figure S5: The (a) observed and (b) predicted distributions of the intermediate species (prevalence = 30%) using noisy absences randomly located or (c) located following a distance gradient from the center of the environmental niche. For each noisy absence type, the top line of 3 maps refers to EM and the bottom line maps to IEM. For each line, noisy absences increase from the left to the right (left 15%, centre 45%, right 75%). The situations with 30 % and 60% of noisy absences are not shown for clarity. The 100 models based on the 100 different learning data sets were used and we evaluated the percentage of mispredicting models in each pixel. The darker the pixels, the higher the percentage of prediction errors.



(M4) : The iterative ensemble modelling approach increases the accuracy of fish distribution models

Christine LAUZERAL[1,2], Gaël GRENOUILLET[1,2] & Sébastien BROSSE[1,2]

[1] Université de Toulouse, UPS, ENFA; UMR5174 EDB (Laboratoire Évolution et Diversité Biologique); 118 route de Narbonne, F-31062 Toulouse, France.

[2] CNRS; UMR5174 EDB, F-31062 Toulouse, France.

Corresponding author: Christine Lauzeral, Laboratoire Evolution et Diversité Biologique, U.M.R 5174, C.N.R.S - Université Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex 4, France. Email: christine.lauzeral@univ-tlse3.fr

ABSTRACT

Species distribution models (SDMs) are widely used to predict the present and future distribution of species. However non environmental absences are known to affect the quality of the models. These absences occur in most databases and are particularly frequent for mobile, cryptic and difficult to detect species, which is the case for freshwater fish. Here, we compared the ability of classical ensemble modelling (EM) and of a new iterative ensemble modelling (IEM) approach, designed to deal with non environmental absences, to predict the distribution over France of 31 fish species. Compared to a classical ensemble modelling approach, IEM improved models accuracy of most of the species having a low detectability. Its performances remained nevertheless limited for some highly detectable species. These results are congruent with those obtained on virtual species in a previous study. They show that the iterative approach is of interest to model the distribution of difficult to detect species, provided that presence data are representative of the potential niche of the species. The IEM approach can also be helpful to predict the potential niche of species under non equilibrium state such as threatened species experiencing a spatial range reduction or non-native species potential invasion extent.

Key-words: Bioclimatic models, niche models, ensemble modelling, species distribution, consensus, freshwater fish.

Short title: Iterative SDMs and fish distribution

INTRODUCTION

Freshwaters are among the most anthropogenicaly threatened ecosystems through habitat destruction, biological invasions, pollution and overexploitation (Butchart, S. H. M. et al. 2010, Wilcove, D. S. et al. 1998). They will also face serious threats through climate changes, as fish dispersion is constrained by the structure of the river networks. Hence, a high proportion of strictly freshwater species will soon be at risk of extinction due to their inability to migrate from one river basin to another (Rahel, F. J. 2007). In Europe, this is particularly true for coldwater fishes that are predicted to lose a large part of their distribution through climate change (Buisson, L. et al. 2008b). Increased efforts are thus needed to identify freshwater species responses to environmental change and to fully use existing data to guide conservation efforts for freshwater ecosystems.

Species distribution models (SDM) are increasingly applied as predictive tools for purposes of conservation planning and management. They are usually based on the use of presence-absences data, but although the presence of a species is factual, absence can have a multiple meaning. Lobo *et al.* (2010) listed three distinct types of absences: environmental absences (the environmental conditions do not allow the presence of the species), contingent absences (the environmental conditions are favorable but other factors such as biotic interactions, barriers to dispersion or local extinction are responsible for the absence of the species) and methodological absences (the species is present but not detected). Contingent absences makes the SDMs predicting the realized niche rather than the potential one, whereas methodological absences are known to reduce the reliability of EM predictions (Lobo, J. M. 2008). Unfortunately, both types of non environmental absences are abundant in fish occurrence databases. On one hand, fish are difficult to detect and their detectability depends on the species, the fishing method and the river size. On the other hand, the web structure of the rivers limits dispersal and anthropogenic events lead to extirpate some species from a part of

their initial niche, or to introduce and assist spread of exotic species that are hence inhabiting only a part of their potential niche. These factors are an important source of non environmental absences.

Presence-only SDMs, by considering only the presence of the species to determine its niche (Farber, O. and Kadmon, R. 2003, Hirzel, A. H. et al. 2002), offer an alternative to the problem of absence uncertainty. However, they frequently overestimate potential distributions (Zaniewski, A. E. et al. 2002), making the prediction of little interest for conservation planning. Binomial likelihood models are specially designed to account for potential sampling errors and have recently been used in the context of global change (Kéry, M. et al. 2010, Moritz, C. et al. 2008, Rowe, R. J. et al. 2010). Unfortunately, these models are designed to be computed using species abundance data or long term survey data that are both rarely available.

Here we tested the iterative ensemble model (hereafter called IEM; Lauzeral et al., summited) approach, designed to reduce the effect of noisy absences, on 31 freshwater fish species. That method has been proved efficient to deal with noisy absences using virtual species, but it remains to be tested on real species distribution data.

In this context, the main objective of this study was to compare the performances of the classical EM and the new IEM in predicting both the spatial distribution of individual fish species and the composition of assemblages. To do this, we used an extended dataset containing climate and physical characteristics and stream fish occurrences in France. We modelled the spatial distribution of 31 fish species in 1110 stream sections, and we compared the values predicted by EM and IEM techniques with observed values obtained from thoroughly surveyed sites.

MATERIALS AND METHODS

Data

An extensive dataset of freshwater fish occurrences covering the entire French territory was used. It was provided by the French Office National de l'Eau et des Milieux Aquatiques (ONEMA) (see Buisson, L. and Grenouillet, G. 2009 for more details). The whole dataset was comprised of a total of more than 10 000 occurrence records for 31 fish species in 1110 stream sections (hereafter referred to as 'sites'). For each site, the environmental characteristics known to be the most relevant descriptors for habitat requirements of fish were measured and only those weakly correlated to each other were retained in this study (see Buisson, L. et al. 2008a for more details). Six environmental descriptors were available for each site: distance from the headwater source (DIS, km); surface area of the drainage basin above the sampling site (SDB, km^2); elevation (m); slope (SLO, ‰); mean stream width (WID, m) and depth (DEP, m). Following Buisson *et al.* (2008a), principal component analysis (PCA) was used to eliminate the colinearity between SDB and DIS. The first axis of the PCA was kept as a synthetic variable describing the longitudinal gradient G. We also constructed the approximation V of local velocity derived from the Chezy formula:

$$V = \log WID + \log DEP + \log SLO - \log(WID + 2DEP)$$

Elevation was log-transformed to meet the hypothesis of normal distribution. The CRU CL 2.0 (Climatic Research Unit Climatology 2.0 version) dataset (New, M. et al. 2002) with a resolution of 10' x 10' was used to describe the current climate. Three climatic variables related to ecological requirements of fish were retained: the mean annual air temperature, the mean annual air temperature range and the mean annual precipitation. Air temperatures were used as a substitute for water temperatures, which are currently not available for all French streams. Indeed, since streams and rivers are reasonably well-mixed water bodies that easily

exchange heat with the atmosphere, it has already been found that air and river water temperatures show a strong positive correlation (e.g., Caissie, D. 2006).

Selecting learning and test data

The whole dataset was first split into two parts, acting as learning and testing subsets. Following Pineda & Lobo (2009), we chose well-monitored sites as test sites to reduce methodological absences. Well-monitored sites were those satisfying the following two criteria: (1) at least 15 successive fish collections performed (1 or 2 times each year) and (2) an average species saturation curve that stabilized at the end of sampling effort. The average saturation curve resulted from a mean value of 100 saturation curves generated by sampling the data randomly. We hence retained as test sites the locations where less than one new species appeared on the average saturation curve during the last 5 occasions the site was fished. This provided a testing subset of 191 sites, which were distributed throughout the country (Fig. 1). The remaining 919 sites were kept as learning data. In the learning data set, fish occurrence came from a single recent (less than 5 years ago) sampling occasion, and hence obviously contained a substantial proportion of methodological absences (i.e., undetected species).

Species detectability

The detectability of each species was evaluated on the 191 well-monitored sites. It was calculated as the ratio between the number of fishing campaigns where the species was detected (cumulated over sites where the species was detected at least once) and the whole number of campaigns on the considered sites (Table 1).

EM and IEM modelling

According to the EM framework, we used six predictive modelling methods: generalized linear models (GLM); generalized additive models (GAM); boosted trees (BT); classification and regression trees (CART); generalized boosted regression models (GBM) and linear discriminant analysis (LDA). For the GLM and LDA models, squared variables were included in the model to deal with non-linearity. The modelling followed the classical EM process (Araújo, M. B. and New, M. 2007). The six statistical methods were used to build a mean probability of presence model (Marmion, M. et al. 2009). The probability vector was converted into a presence-absence response using a cut-off threshold determined by maximizing the Kappa index.

In the IEM procedure, the data matrix predicted using the EM framework was used to update the raw data set. We considered noisy absences to be the false presences predicted by the EM model (i.e., the cases where the model predicted species presence while it was actually absent from the training set). These noisy absences were then considered as presence and the resulting new data matrix was used as a new model training set. This post-processing of model outputs was iterated until predictions stabilized, therefore providing a potential distribution of the species (see Lauzeral *et al*, summited for more detail). The entire procedure was repeated 150 times. The modelling procedure was implemented in R (R Development Core Team 2011).

To evaluate the prediction variability inherent to the statistical methods (i.e., GBM and BT), we ran the EM 150 times for each species. We observed that the maximum number of sites with variable predictions was reached in less than 30 runs and that less than 3% of the sites had variable predictions. We thus considered that our IEM model had stabilized when less than 3% of the sites provided variable predictions in 30 successive iterations.

The evolution of the variability among the six SDM predictions through the iterative process was evaluated at each iteration. Following Thuiller (2004), we performed a standardized Principal Component Analysis (PCA) on the data matrix made of the 6 probability-of-presence vectors at the 191 test sites, and we evaluated the consensus among the predictions by calculating the percentage of variance accounted for by the first axis of the PCA.

Comparing IEM and EM

The IEM was run on the 31 fish species and the results obtained after the 150 iterations were compared to those obtained using the classical EM (i.e., those obtained at the end of the first IEM iteration).

For each of the 31 fish species, we first evaluated the predictive accuracy of both EM and IEM on the 191 test sites by measuring one threshold independent (AUC) and three threshold dependant indicators: the percentage of mispredicted sites (i.e., both false absences and false presences); the Kappa index; and the half-sum of sensitivity and specificity.

Then, we determined the predicted species richness per site by summing occurrence predictions of the 31 species. The efficiencies of EM and IEM in predicting species richness were compared by assessing the relationship between observed and predicted species richness on the 191 test sites. Finally, we compared the predicted species assemblages to the observed assemblages by calculating a Jaccard similarity index between observed and predicted assemblages. Wilcoxon tests were used to make pairwise comparisons between the two modelling methods.

RESULTS

Prevalence in the learning data set was highly variable among fish species, ranging from 0.036 for burbot (*Lota lota*) to 0.55 for stone loach (*Barbatula barbatula*). Detectability also showed high variability (i.e., 8-fold variation among species) and ranged from 20.1% (*Cyprinus carpio*) to 89.4% (*Barbus meridionalis*) (Table 1). This large variation in species detectability suggested large variations in the occurrence of noisy absences in the learning data set.

IEM modelling

For all 31 species, the iterative process tended to converge rapidly, as the predictions stabilized after 2 to 35 iterations, depending on the species. It should however be noticed that 2 species (i.e., bleak, *Alburnus alburnus* and gudgeon, *Gobio gobio*) showed a temporary stabilization (less than 6 sites had variable predictions during more than 10 iterations), then predictions became variable again before they stabilized permanently. For these species, the final predictions differed from intermediate stable predictions for less than 5% of the sites.

After a few iterations, the 6 different methods provided consensual predictions for the 191 test sites. At the first iteration (i.e., the EM), the mean percentage of variance accounted for by the first axis of the PCA was 85.6%. Using IEM, consensus increased after 15 iterations up to 93.2%, and then reached a plateau up to the end of the iterative procedure.

Predictive performances

At the species level, compared to EM, IEM significantly reduced false absences (Wilcoxon test, $p<0.001$). Due to the IEM principle (i.e., replacing noisy absences by presences), false presences increased significantly (Wilcoxon test, $p<0.001$, Fig. 2). Lowering false absences and increasing false presences led to a percentage of errors that did not differ significantly

between IEM and EM (Wilcoxon test, p=0.18). However, the two other threshold dependent indices showed that IEM performed better than EM (Kappa: Wilcoxon test, p<0.01, TSS: Wilcoxon test, p<0.001) (Fig. 3). In particular, the Kappa index calculated for IEM testified for a good score (>0.6) for 12 species and a moderate score (between 0.4 and 0.6) for 17 species. Our predictions were thus reliable for 29 out of the 31 species (i.e., 94% of the species). EM performance was clearly lower with only 8 species reaching a Kappa score above 0.6 and just 23 species for which the predictions were reliable (i.e., 74% of the species). The AUC showed a significant (p<0.01) but limited decrease (Fig. 3). Species that benefited most from iterations were some of the rare ones (nase, *Parachondrostoma nasus*; salmon, *Salmo salar* and stickleback, *Gasterosteus aculeatus*). Species that did not benefit from iterations were some of the most common (stone loach, *Barbatula barbatula* and gudgeon, *G. gobio*). More generally, species benefiting from iterations were the less detectable ones: the variation of the Kappa index and of the TSS significantly decreased as detectability increased (p<0.05 and p<0.01 respectively) (Fig. 4).

At the assemblage level, the relationship between observed and predicted richness was highly significant for EM (r^2=0.90; p<0.001, Fig. 5a) which reliably predicted species richness in sites containing few species. However, it tended to underestimate the species richness of the richest sites as it predicted on average 68% of the observed richness. This bias was reduced using IEM as the predicted species richness was on average 89% of the observed one, with a highly significant relationship between observed and predicted values (r^2=0.91; p<0.001; Fig. 5b). Finally, IEM also increased the similarity (i.e., Jaccard index) between observed and predicted fish assemblages from 0.45 ± 0.25 to 0.50 ± 0.26 (Wilcoxon test, p<0.001).

DISCUSSION

Inclusion of reliable absences within the calibration dataset, by giving information on the lack of suitability of some places, significantly improved model predictions (Engler, R. et al. 2004, Lobo, J. M. 2008). Unfortunately, absences have multiple sources: environmental, contingent and methodological. Absence is thus the main cause of uncertainty in species occurrence data matrices. This can have detrimental consequences on the relevance of presence-absence SDMs (Lobo, J. M. 2008, Lobo, J. M. et al. 2010), especially for fish as both types of non environmental absences are abundant in fish occurrence databases.

By dealing with methodological absences, the IEM provided a way to consider almost all the species whatever their detectability, as the spatial distribution of most of the 31 considered species was better predicted using IEM than using EM. These results parallel and extend those of Lauzeral et al. (summitted) on virtual species. This testifies that the IEM method is transferable to real species having obviously more complex niches than virtual niches simulated using a reduced set of environmental variables. Some differences in prediction performance should however be noticed according to the species. Although easily detectable species benefited less from iterations, IEM still provided better predictions than EM for most of these species. Only 2 out of the 12 highly detectable species (i.e. detectability > 0.6; gudgeon, *G. gobio* and stone loach, *B. barbatula*) were better predicted using EM than using IEM. This was probably due to the high prevalence of these species (present in 46.5% and 55.2% of the learning data sites, respectively), since when a species is widely present, the iterative process maximizes presences and hence tends to increase the number of false presences. As the number of non-colonized sites is low, specificity is sensitive to a small increase in false presences, so specificity may decrease faster than sensitivity increases, thus the sum of these two indices decreases. Such a result parallels previous studies showing that wide niches (i.e., generalist species) can decrease model quality (Hernandez, P. A. et al. 2006,

Kadmon, R. et al. 2003). Besides, the accuracy of the predictions markedly increased using IEM compared to EM for species difficult to detect as for most of the species having a low prevalence. The distribution of these species is recognized as difficult to predict using EM, as low detectability and low prevalence generally decrease the consensus between modelling methods (Pearson, R. G. et al. 2007, Wisz, M. S. et al. 2008). Among 10 poorly detectable species (i.e. detectability < 0.4), only two (black bullhead, *Ameiurus melas* and ruffe, *Gymnocephalus cernuus*) did not markedly benefit from iterations. The first one exhibited a small decrease of Kappa and an increase of TSS. The second one showed no change in model quality. These two species are represented by only a few occurrences in our data (Table 1), and too much ecological information was probably missing to be able to properly model species niche. Moreover, the black bullhead is an invasive species originating from North America that shows strong temporal fluctuations in local distribution due to both rapid colonization abilities combined to massive local population declines due to bacterial and viral diseases (Keith, P. and Allardi, J. 2001). This probably makes the black bullhead niche particularly complex to model with a limited set of occurrences. In the same way, for the ruffe, none of the presence sites in the learning dataset was located in the western part of France (characterized by oceanic climate) while a fourth of the presence test sites for this species were located in this area. Again, a large part of the environmental niche was lacking from the learning data set and IEM was not able to determine the entire species niche and hence to fill false absences. For these two species, the percentage of noisy absences falls over a critical percentage making the IEM unable to fill the gaps in the dataset as it has also been observed on virtual species (Lauzeral *et al.*, summited). Apart from these particular species, fish with low prevalence and low detectability, such as burbot (*Lota lota*), bitterling (*Rhodeus amarus*) or rudd (*Scardinius erythrophthalmus*), were better predicted using IEM. This testifies that IEM can predict the distribution of rare species that are usually excluded from

the predictive modelling approaches especially when using presence-absence models (Pearson, R. G. et al. 2007, Wisz, M. S. et al. 2008).

Such a characteristic makes the IEM useful to predict the potential distribution of threatened species. Indeed, threatened species have often been extirpated from a large part of their suitable area, and hence have a low prevalence in the datasets with a lot of their absences being contingent, making their potential distribution difficult to predict using classical SDMs (Cianfrani, C. et al. 2010). In the same way, IEM might also be efficient for the early warning of invasive species spread. It is indeed recognised that most non-native species are in a non-equilibrium state, particularly due to a differential propagule pressure and human impact on ecosystems across the world (Blanchet, S. et al. 2009, Leprieur, F. et al. 2008). The way currently proposed to predict spatial invasion range of invasive species is the calibration of models on the niche conditions found in both the native and the exotic range of the species (e.g., Beaumont, L. J. et al. 2009), aiming to account for potential niche shifts between native and invasion ranges (Beaumont, L. J. et al. 2009, Lauzeral, C. et al. 2011, Medley, K. A. 2010, Rödder, D. and Lötters, S. 2009). This however strongly limits the predictive efficiency of the models, as a substantial part of the absences in the exotic range are contingent. The IEM hence constitutes an alternative for predicting the invasion potential of current and future invaders as soon as both data on native and invasive ranges are available.

As IEM is poorly sensitive to species prevalence and species detectability, it can also be used to provide predictions in a multi-species context and hence extend our predictions from species to assemblages. Although EM reliably predicted the species richness in the sites containing few species, it underpredicted the richness of rich assemblages, in which the detectability of some species is low (Kéry, M. and Schmid, A. 2006). These last species increase the percentage of methodological absences and thus decrease the accuracy of models' predictions, leading to the prediction of unrealistic assemblages. IEM provides a

more realistic prediction of both the observed richness and the assemblage composition, whatever the species richness of the sites. This corroborates the idea that the IEM is efficient in filling methodological absences.

Even if the IEM has been proved to increase the SDMs accuracy of both virtual and real species as soon as the data set contains abundant non environmental absences, it remains to be evaluated in more details before being intensively used. Its sensitivity to parameters known to affect SDMs outputs (cut-off threshold selection, species prevalence…), to inaccurate environmental variables but also to false absences remains to be assessed.

Acknowledgements

This study was supported by the BIOFRESH European project (FP7-ENV-2008). We are indebted to the Office National de l'Eau et des Milieux Aquatiques (ONEMA) for providing fish data.

References

Araújo, M. B. and New, M. 2007. Ensemble forecasting of species distributions. — Trends in Ecology and Evolution 22: 42-47.

Beaumont, L. J. et al. 2009. Different climatic envelopes among invasive populations may lead to underestimations of current and future biological invasions. — Diversity and Distributions 15: 409-420.

Blanchet, S. et al. 2009. Broad-scale determinants of non-native fish species richness are context-dependent. — Proceedings of the Royal Society B, Biological Sciences 276: 2385-2394.

Buisson, L. et al. 2008a. Modelling stream fish species distribution in a river network: the relative effects of temperature versus physical factors. — Ecology of Freshwater Fish 17: 244-257.

Buisson, L. and Grenouillet, G. 2009. Contrasted impacts of climate change on stream fish assemblages along an environmental gradient. — Diversity and Distributions 15: 613-626.

Buisson, L. et al. 2008b. Climate change hastens the turnover of stream fish assemblages. — Global Change Biology 14: 2232-2248.

Butchart, S. H. M. et al. 2010. Global Biodiversity: Indicators of Recent Declines. — Science 328: 1164-1168.

Caissie, D. 2006. The thermal regime of rivers: a review. — Freshwater Biology 51: 1389-1406.

Cianfrani, C. et al. 2010. Do habitat suitability models reliably predict the recovery areas of threatened species? — Journal of Applied Ecology 47: 421-430.

Engler, R. et al. 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. — Journal of Applied Ecology 41: 263-274.

Farber, O. and Kadmon, R. 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the Mahalanobis distance. — Ecological Modelling 160: 115-130.

Hernandez, P. A. et al. 2006. The effect of sample size and species characteristics on performance of different species distribution modeling methods. — Ecography 29: 773-785.

Hirzel, A. H. et al. 2002. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? — Ecology 83: 2027-2036.

Kadmon, R. et al. 2003. A systematic analysis of factors affecting the performance of climatic envelope models. — Ecological Applications 13: 853-867.

Keith, P. and Allardi, J. 2001. Atlas des poissons d'eau douce de France. — Muséum National d'Histoire Naturelle.

Kéry, M. et al. 2010. Predicting species distributions from checklist data using site-occupancy models. — Journal of Biogeography 37: 1851-1862.

Kéry, M. and Schmid, A. 2006. Estimating species richness: calibrating a large avian monitoring programme. — Journal of Applied Ecology 43: 101-110.

Lauzeral, C. et al. 2011. Identifying climatic niche shifts using coarse-grained occurrence data: a test with non-native freshwater fish. — Global Ecology and Biogeography 20: 407-414.

Leprieur, F. et al. 2008. Fish invasions in the world's river systems: when natural processes are blurred by human activities —PloS Biology 6: e28. http://dx.doi.org/10.1371/journal.pbio.0060028.

Lobo, J. M. 2008. More complex distribution models or more representative data? — Biodiversity Informatics 5: 14-19.

Lobo, J. M. et al. 2010. The uncertain nature of absences and their importance in species distribution modelling. — Ecography 33: 103-114.

Marmion, M. et al. 2009. Evaluation of consensus methods in predictive species distribution modelling. — Diversity and Distributions 15: 59-69.

Medley, K. A. 2010. Niche shifts during the global invasion of the Asian tiger mosquito, *Aedes albopictus* Skuse (Culicidae), revealed by reciprocal distribution models. — Global Ecology and Biogeography 19: 122-133.

Moritz, C. et al. 2008. Impact of a century of climate change on small-mammal communities in Yosemite National Park, USA. — Science 322: 261-264.

New, M. et al. 2002. A high-resolution dataset of surface climate over global land areas. — Climate Research 21: 1-25.

Pearson, R. G. et al. 2007. Predicting species distributions from small numbers of occurrence records: a test case using cryptic geckos in Madagascar. — Journal of Biogeography 34: 102-117.

Pineda, E. and Lobo, J. M. 2009. Assessing the accuracy of species distribution models to predict amphibian species richness patterns. — Journal of Animal Ecology 78: 182-190.

R Development Core Team 2011. R: A language and environment for statistical computing. — R Foundation for Statistical Computing.

Rahel, F. J. 2007. Biogeographic barriers, connectivity and homogenization of freshwater faunas: it's a small world after all. — Freshwater Biology 52: 696-710.

Rödder, D. and Lötters, S. 2009. Niche shift versus niche conservatism? Climatic characteristics of the native and invasive ranges of the Mediterranean house gecko (*Hemidactylus turcicus*). — Global Ecology and Biogeography 18: 674-687.

Rowe, R. J. et al. 2010. Range dynamics of small mammals along an elevational gradient over an 80-year interval. — Global Change Biology 16: 2930-2943.

Thuiller, W. 2004. Patterns and uncertainties of species' range shifts under climate change. — Global Change Biology 10: 2020–2027.

Wilcove, D. S. et al. 1998. Quantifying threats to imperiled species in the United States. — Bioscience 48: 607-615.

Wisz, M. S. et al. 2008. Effects of sample size on the performance of species distribution models. — Diversity and Distributions 14: 763-773.

Zaniewski, A. E. et al. 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. — Ecological Modelling 157: 261-280.

Table 1: Fish species prevalence in the learning (n=919) and test (n=191) data set and detectability of the species. The detectability is the ratio between the number of sampling runs when the species is detected (cumulated over sites) and the whole number of sampling runs at the selected sites.

Code	Species name	Learning prevalence	Test prevalence	Detectability
Abb	*Abramis brama*	75	41	39.9
Ala	*Alburnus alburnus*	160	56	66.7
Alb	*Alburnoides bipunctatus*	107	32	43.2
Amm	*Ameiurus melas*	48	17	39.4
Ana	*Anguilla anguilla*	310	98	81.2
Bab	*Barbatula barbatula*	508	140	77.8
Bam	*Barbus meridionalis*	58	9	89.4
Bar	*Barbus barbus*	177	49	60.2
Blb	*Blicca bjoerkna*	59	36	44.2
Cog	*Cottus gobio*	446	129	78.3
Cyc	*Cyprinus carpio*	58	38	20.1
Esl	*Esox lucius*	149	66	57.2
Gaa	*Gasterosteus aculeatus*	88	49	28.3
Gog	*Gobio gobio*	446	126	74.1
Gyc	*Gymnocephalus cernuus*	44	32	38.6
Lap	*Lampetra planeri*	218	94	59.1
Leg	*Lepomis gibbosus*	121	61	41.7
Lel	*Leuciscus leuciscus*	214	71	57.8
Les	*Leuciscus souffia*	61	16	74.5
Lol	*Lota lota*	33	14	30.9
Pan	*Parachondrostoma nasus*	73	32	43.2
Pat	*Parachondrostoma toxostoma*	40	9	55.1
Pef	*Perca fluviatilis*	227	85	64.8
Php	*Phoxinus phoxinus*	502	151	75.8
Pup	*Pungitius pungitius*	64	37	39.9
Rha	*Rhodeus amarus*	68	27	39.5
Rur	*Rutilus rutilus*	332	102	69.9
Sas	*Salmo salar*	37	41	47.9
Sce	*Scardinius erythrophtalmus*	87	65	32.7
Sqc	*Squalius cephalus*	418	105	77.3
Tit	*Tinca tinca*	111	66	37.6

Figures:

Figure 1: Geographical distribution of the 919 fish sampling sites (white dots) used as learning dataset, and the 191 sites (black dots) identified as well-monitored sites and used as testing dataset (see Material and Methods).

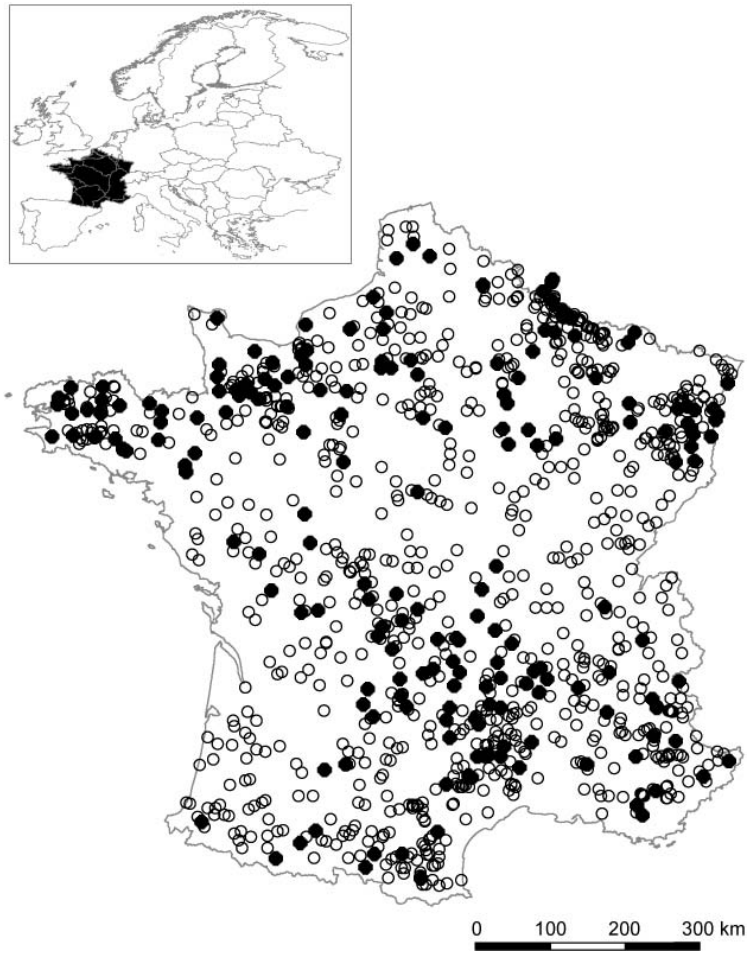


Figure 2: Proportion of the two types of mispredicted sites across the 31 fish species after the first iteration (EM, white) and at the end of the process (IEM, grey). FA: false absence (i.e., the species was detected but was predicted as absent); FP: false presence (i.e., the species was not detected but was predicted as present).

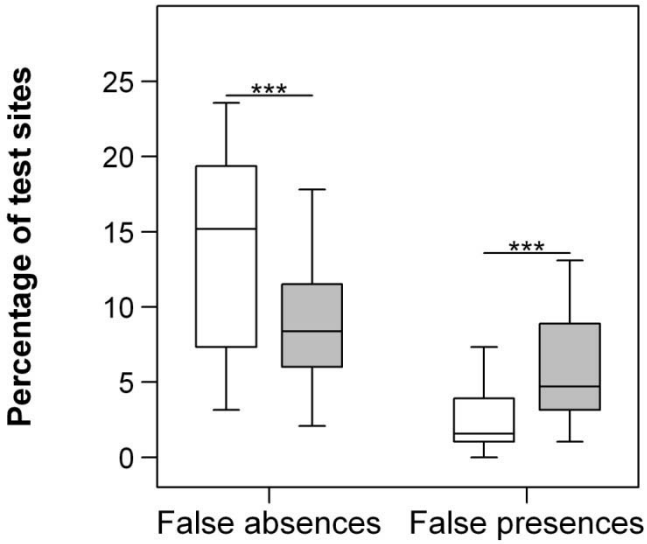


Figure 3: Model evaluation of the two SDMs, EM (white) and IEM (grey), using a threshold independent measure (AUC) and two threshold dependant measures (Kappa and TSS).

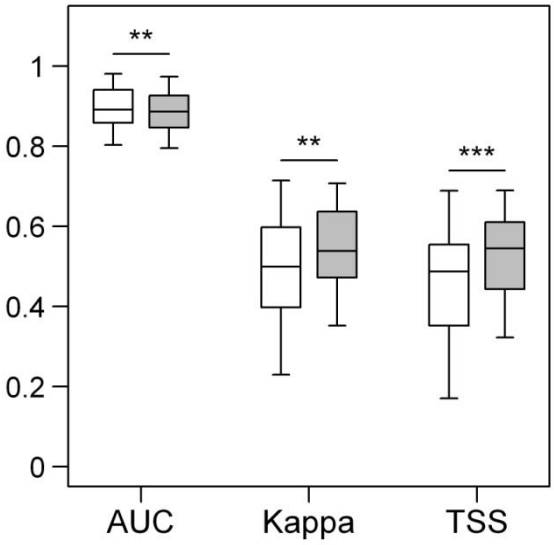


Figure 4: Variation of quality indices between EM and IEM for the 31 species according to the species detectability calculated over the 191 test sites (see Material and Methods): a) Kappa index; b) TSS. The size of the dots is proportional to the prevalence of the species in the learning data set. Species codes as in Table 1.

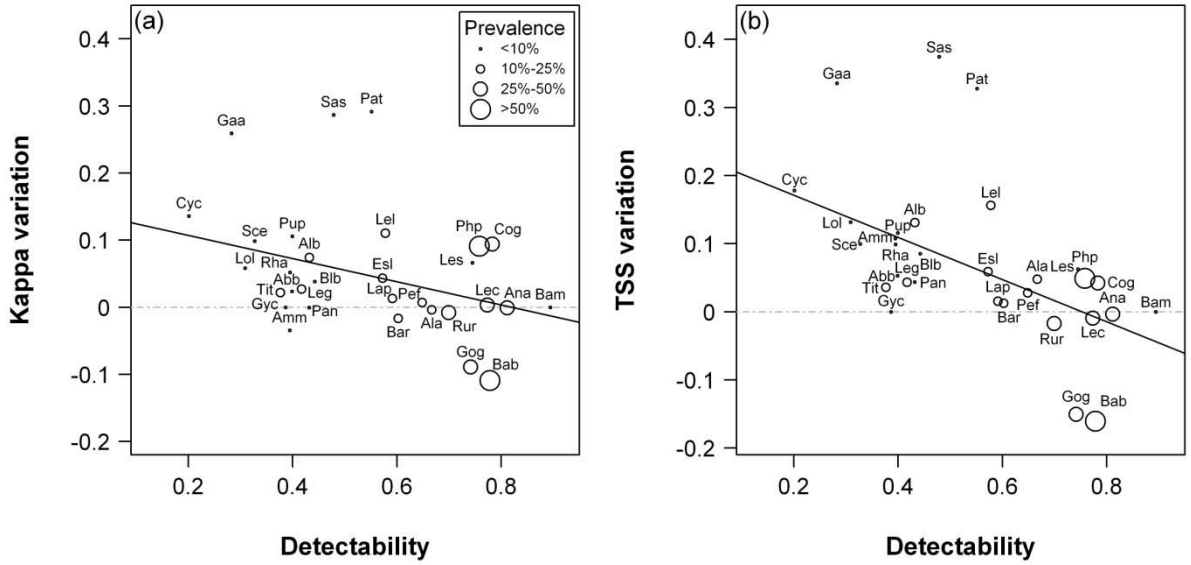
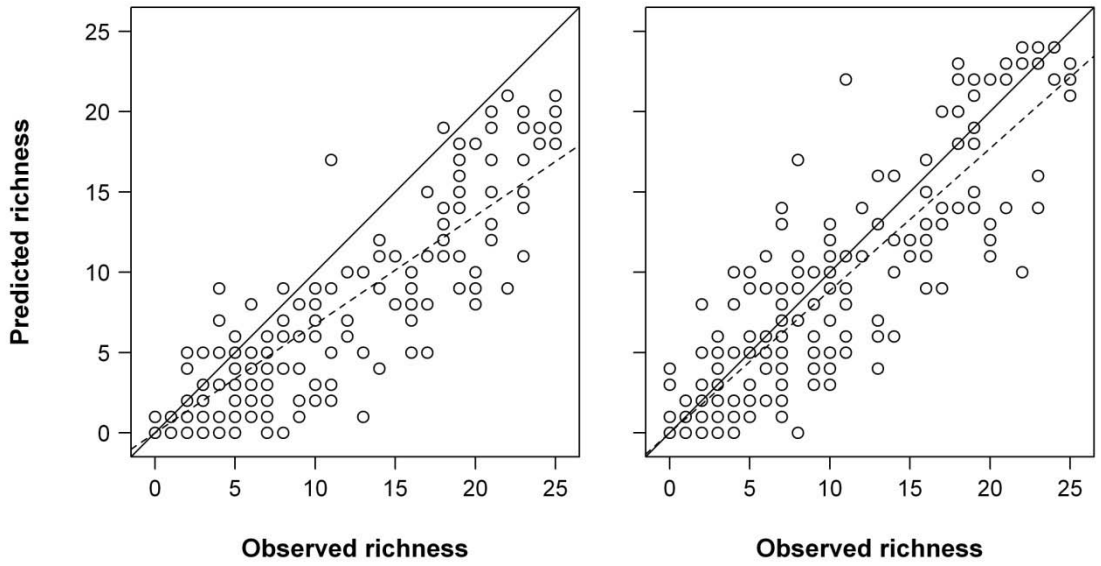


Figure 5: Relationship between observed and predicted fish species richness in the 191 test sites (a) EM; b) IEM). The dashed line represents the linear relationship between observed and predicted richness (y=0.68x r^2=0.90, p<0.001 for EM; y=0.89x, r^2=0.91, p<0.001 for IEM). The solid line represents the perfect fit line (y=x).



Annexe

SPRICH: a database of freshwater fish species richness across the World

Sébastien Brosse, Olivier Beauchard, Simon Blanchet, Hans H. Dürr, Gaël Grenouillet, Bernard Hugueny, Christine Lauzeral, Fabien Leprieur, Pablo A. Tedesco, Sébastien Villéger, Thierry Oberdorff

S. Brosse (corresponding author), G. Grenouillet, C. Lauzeral, S. Villeger

Laboratoire Evolution et Diversité Biologique, UMR 5174, Université Paul Sabatier - CNRS -ENFA, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France.

Email: sebastien.brosse@univ-tlse3.fr ; Phone: 33 5 61 55 67 47

O. Beauchard

Department of Biology, Faculty of Sciences, Ecosystem Management Research Group, University of Antwerp, Universiteitsplein 1, BE-2610 Antwerpen (Wilrijk), Belgium

S. Blanchet

Station d'Ecologie Expérimentale du CNRS à Moulis, U.S.R 2936, 09200 Moulis, France

H.H. Dürr

Department of Physical Geography, Faculty of Geosciences, Heidelberglaan 2, P.O. box 80.115, Room 106 Utrecht University, NL-3508 TC Utrecht, The Netherlands.

F. Leprieur

Laboratoire Ecologie des Systèmes Marins Côtiers, U.M.R. 5119, CNRS-IFREMER-UM2-IRD- Université Montpellier 2, Place Eugène Bataillon, F-34095 Montpellier Cedex 5, France.

B. Hugueny, P.A. Tedesco, T. Oberdorff

UMR ''BOREA'' CNRS 7208/IRD 207/MNHN/UPMC, DMPA, Museum National d'Histoire Naturelle, 43 rue Cuvier, 75231 Paris Cedex, France .

Abstract

A growing interest is devoted to large-scale approaches in ecology, for both plants and animals. In particular, macroecological studies allow examination of the patterns and determinants of species richness of various organisms across the world, which might have important implications in the prediction and the mitigation of the consequences of global change. Here, we provide richness data for freshwater fish, which are the most diverse vertebrate taxa with more than 13 000 species living in freshwater described to date. We conducted an extensive literature survey of native, non-native and endemic freshwater fish species richness and the resulting database, called SPRICH, was gathered from more than 400 bibliographic sources including published papers, books and grey literature databases. SPRICH contains richness values at the river basin scale for 1054 river basins covering more than 80% of the earth's continental surface. This database is currently the most comprehensive global database for native, exotic and endemic freshwater fish richness at the river basin scale.

Key words

Global extent, River drainage basin, Fishes, Native, Endemic, Exotic.

Introduction

The emergence of macroecology has led to an increase in the number of studies examining the patterns and determinants of species richness of various organisms across the world (Brown, 1995; Gaston & Blackburn, 2000). Besides improving theoretical knowledge, macroecological approaches might have important implications in the prediction and the mitigation of the consequences of global change for organisms and ecosystem functions (Kerr et al., 2007). Macroecological studies require global scale datasets that actually exist only for a few animal taxa, namely freshwater fish (Leprieur et al., 2008), birds (Davies et al., 2007), amphibians and mammals (Orme et al., 2005; Grenyer et al., 2006). Concerning freshwater fish, since the nineties, several studies investigated the continental patterns of fish species richness in west African rivers (Hugueny, 1989) as well as in European and north American rivers (Oberdorff et al., 1997). Merging those data allowed to build up a global database containing species richness information on c.a. 200 river basins, and to develop for the first time global scale approaches at the river basin grain. This permitted to investigate global richness determinants of freshwater fish (Oberdorff et al., 1995; Guegan et al., 1998), but still not permitted to draw a global map of fish richness due to an insufficient coverage of the world continental areas. This original data was then amplified by an extensive literature survey to reach more than 1000 basins throughout the world, giving rise of the SPRICH database that covers more than 80% of the earth continental surface. SPRICH database hence permitted to accurately map global species richness patterns of native (Oberdorff et al., 2011), endemic (Tedesco et al., 2012) and non-native (Leprieur et al., 2008) fish species richness. The corresponding raw data has however never been published to date. As such a large scale data can be of interest for most freshwater ecologists studying species richness patterns, and could also help supporting global decision-making (Seys et al., 2004; Kerr et al., 2007), we here provide the SPRICH database. It contains species richness for native, non-native and endemic freshwater fish in 1054 basins dispersed throughout the world, as well as location, area and altitudinal range of each basin.

Data collection and availability of the database

Data collection lasted from 2003 to 2008, as a joint collaboration between two French research institutes: the University Paul Sabatier, Toulouse (S. Brosse, O. Beauchard, F. Leprieur and S. Blanchet) and the National Museum of Natural History (MNHN) in Paris (T. Oberdorff and P.A. Tedesco). During this period, we conducted an extensive survey of literature published from 1960 to 2008 on native, exotic and endemic freshwater fish species richness at the river basin grain. Only complete richness counts at the river basin grain

were considered. We discarded incomplete species counts or check lists such as local inventories of a stream reach or inventories based solely on a given family, but we used these partial data for cross-checking available species counts at the basin scale. Only freshwater fish were considered. Marine species occasionally occurring in freshwater and estuarine species with no freshwater life stage were discarded from our richness counts. The resulting database was gathered from more than 400 bibliographic sources including published papers, books, grey literature and web-based sources. The complete list of references used to build up the SPRICH database is given as an online supplementary material (ESM 1) and the original bibliography is stored at University Paul Sabatier, Toulouse, France and at the National Museum of Natural History (MNHN), Paris, France.

Three species richness counts were made for each basin: native, exotic and endemic richness.

Native richness is the number of species that historically occur in the basin. It account for both endemic and non-endemic species, but excludes exotic species that have been directly or indirectly (e.g. via artificial channels) introduced in the basin. It also excludes extirpated species that are currently absent in the basin. Exotic richness is the number of established non-native species occurring in each basin. We considered as non-native a species (i) that did not historically occur in a given basin and (ii) that successfully established, i.e., self-reproducing populations. Endemic richness only refers to narrow endemics that are native from a single river basin (i.e. single-drainage endemics as the analogy of “single-island endemics” often applied in insular systems).

The location of each basin is informed by the latitude and longitude at the river mouth, minimal, maximal and median latitude and longitude values of the watershed. Latitude and longitude values have been collected from the literature and from world atlases (Encartra and Google earth). For each basin we indicated minimal and maximal altitude as well as basin area. When this information was not available in the literature, it was inferred from global atlases.

Data collection was completed in 2008 and data entries were carefully reviewed by the project contributors. The SPRICH data is given as an online supplementary material to this article (ESM 2). The database is hence freely available pending citation of the present paper.

The SPRICH content

The information given in the SPRICH database (ESM 2) is organized vertically by river basin sorted alphabetically. Fish richness and environmental descriptors are listed horizontally. Columns are as follows:

- *Basin* is the river basin name as found in the literature. When a basin was unnamed, it was identified by the name of the country or of the region followed by ".un".

- *TotalR* is the total richness, i.e. the total number of freshwater fish species living in the river basin. It accounts for both native and exotic species.

- *NativeR* is the native richness, i.e. the number of freshwater fish species native to the basin (including endemic species).

- *ExoticR* is the exotic (or non-native) richness, i.e. the number of freshwater fish species introduced into a basin and that have become established (i.e. self-reproducing populations) in the basin.

- *EndemicR* is the endemic richness, i.e. the number of endemic freshwater fish species in the basin. Endemic species are those inhabiting only one river drainage basin (i.e. single-drainage endemics).

- *Brealm* is the biogeographic realm to which the basin belongs, as described by Leveque et al. (2008).

- *Latrivermouth* and *Lonrivermouth* are the geographical coordinates (latitude and longitude) of the river mouth in degrees.

- *Lonmin*, *Lonmax*, *Lonmed*, *Latmin*, *Latmax*, *Latmed* are the minimal, maximal and median coordinates (longitude and latitude in degrees) of the river basin.

- *Minalt*, *Maxalt* are the minimum and maximum altitude in meters above sea level of the basin. Note that the minimum altitude can differ from zero for endorheic basins.

- *Area* is the total surface area covered by the river basin in square kilometers.

Discussion

The SPRICH database contains fish richness values for 1054 river basins covering more than 80% of the earth's continental surface (Fig. 1). This database is currently the most comprehensive global database for native, exotic and endemic freshwater fish richness at the river basin grain. Considering the number of river basins per biogeographic realm reveals that the available information is not balanced throughout the world. While fish species richness of temperate basins of the Northern hemisphere are easily available, it was more difficult to collect relevant information for tropical areas. This was particularly true for the Oriental region where we were not able to gather information on more than 59 basins (Table 1). Such a discrepancy between temperate and

tropical regions is well known in ecology (Jackson & Sweeney, 1995; Blanchet et al., 2009), and underlines the urgent need for more studies in the tropical areas (e.g. Tedesco et al., 2005) that host the highest biodiversity but are also facing strong human disturbances (MEA, 2005; Dudgeon et al., 2006). Nevertheless, basin areas and altitudinal ranges were well balanced despite natural variations explained by the presence of high mountain ranges in some realms, like the Himalaya in the Oriental realm or the Andes in the Neotropics. In addition, the high variability of altitudinal range and basin area within realms testifies that the database is not biased towards a certain basin size in a realm compared to another, avoiding therefore a potential sampling bias between realms.

The overall richness patterns derived from SPRICH are consistent with those found in the literature. The average richness was maximal in the Afrotropical, Neotropical and Oriental realms (c.a. 60 species per basin, Table 1), which are known to host the highest fish diversity with 3000 to 4500 species per realm (Leveque et al., 2008). In the same way, the average number of species per basin was minimal for the Australian realm (c.a. 17 species per basin, Table 1) which also hosts the lowest overall number of species with only 580 species (Leveque et al., 2008). Finally, the average exotic and endemic species richness per basin was also consistent to that found in the literature, with a high rate of endemism in tropical regions and a high rate of invasion in highly anthropized realms (Moyle & Cech, 2004; Leveque et al., 2008).

Although the SPRICH database provides an overall realistic image of fish richness patterns across the world, potential bias could affect the data. First, the basin richness given in the SPRICH data is derived from data published between 1960 and 2008. Recent species descriptions or taxonomic revisions might hence affect the number of species. This might for example be the case for European fish data where the number of European freshwater fish species varies between 256 (Maitland, 2000) and 546 species (Kottelat & Freyhoff, 2007) according to the sources. This two-fold increase is however mainly due to the splitting of some widely distributed species into several local species, leading to slight changes in terms of basin richness (Brosse & Blanchet, unpubl. data). Second, recent extirpations or invasions might affect richness patterns. As shown by Villeger et al. (2011), extirpations remain rare compared to invasions. As a consequence, exotic richness might be the most affected by recent non-native species establishment, which in turn might increase the total richness count. To deal with these potential biases, a permanent update of the database might be of interest. We should however be aware that increasing the data collection time span can also introduce biases due to the high variability of ecological inventories updates across the world. Indeed, richness counts from basins located in

developing countries and/or remote areas often arise from single studies that can difficultly be updated or cross-checked due to the lack of recent or confirmatory literature.

Up to now, the SPRICH database permitted to investigate the global patterns of freshwater fish species richness across the world and to quantify the relative role of basin area, energy and historical processes in determining the fish species richness at the basin grain (Oberdorff et al., 2011). A similar approach focused on endemic richness revealed the spatial patterns and determinants of endemic species richness over the world (Tedesco et al., 2012). Focusing on exotic richness also permitted to demonstrate that human activities, by increasing the species introduction rate and by disturbing the environment, is the main driver of the exotic species establishment across the world, explaining therefore why economically developed areas host the highest richness in exotic freshwater fish (Leprieur et al., 2008; Blanchet et al., 2009).

Apart from these global scale studies, we hope the SPRICH database will open new research opportunities, such as cross-taxa congruence analyses in global richness patterns. In the same way, the SPRICH could be of interest for regional studies on fish diversity, such as biodiversity gradient approaches or species-area investigations. Such macroecological works might have important implications in the prediction and the mitigation of the consequences of global changes (Kerr et al., 2007). Freshwater ecosystems being currently recognized as the most impacted ecosystems on earth (MEA, 2005; Dudgeon et al., 2006), we hope that the SPRICH data will help to develop global conservation programs and contribute to large scale aquatic ecosystems management.

Acknowledgements

This work was supported by the National Agency for Research (ANR) Freshwater Fish Diversity (ANR-06-BDIV-010) and by the EU BIOFRESH project (7th Framework European program, Contract N°226874). EDB is part of the "Laboratoire d'Excellence" (LABEX) entitled TULIP (ANR -10-LABX-41).

Electronic supplementary material

ESM 1. Literature used to build-up the SPRICH database. The full reference of the bibliographic sources is given for each river basin.

ESM 2. The SPRICH database.

References

Blanchet, S., F. Leprieur, O. Beauchard, J. Staes, T. Oberdorff & S. Brosse, 2009. Broad-scale determinants of non-native fish species richness are context-dependent. Proceedings of the Royal Society B 276: 2385-2394.

Brown, J. H., 1995. Macroecology. University of Chicago Press, Chicago.

Davies, R. G., C. D. L. Orme, D. Storch, V. A. Olson, G. H. Thomas, S. G. Ross, T.-S. Ding, P. C. Rasmussen, P. M. Bennett, I. P. F. Owens, T. M. Blackburn & K. J. Gaston, 2007. Topography, energy and the global distribution of bird species richness. Proceedings of the Royal Society B 274: 1189-1197.

Dudgeon, D., A. H. Arthington, M. O. Gessner, Z. I. Kawabata, D. J. Knowler, C. Leveque, R. J. Naiman, A. H. Prieur-Richard, D. Soto, M. L. J. Stiassny, et al., 2006. Freshwater biodiversity: importance, threats, status and conservation challenges. Biological Reviews 81: 163-182.

Gaston, K. J. & T.M. Blackburn, 2000. Pattern and Process in Macroecology. Blackwell Science, Oxford.

Grenyer, R., C. D. L. Orme, S. F. Jackson, G. H. Thomas, R. G. Davies, T. J. Davies, K. E. Jones, V. A. Olson, R. S. Ridgely, P. C. Rasmussen, T. S. Ding, P. M. Bennett, T. M. Blackburn, K. J. Gaston, J. L. Gittleman & I. P. F. Owens, 2006. Global distribution and conservation of rare and threatened vertebrates. Nature 444: 93–96.

Guegan, J.F., S. Lek & T. Oberdorff, 1998. Energy availability and habitat heterogeneity predict global riverine fish diversity. Nature 391: 382–384.

Harrison, I. J. & M. L. J. Stiassny, 1999. The quiet crisis: A preliminary listing of freshwater fishes of the World that are either extinct or ‘‘missing in action’’. In MacPhee, R. D. E. (ed.), Extinctions in Near Time: Causes, Contexts, and Consequences. Plenum Press, New York and London, 271–331.

Hugueny, B., 1989. West African rivers as biogeographic islands: species richness of fish communities. Oecologia 79: 235–243.

Jackson, J. K. & B.W. Sweeney, 1995. Present status and future directions of tropical stream research. Jounal of the North American Benthological Society 14: 5–11.

Kerr, J. T., H. M. Kharouba & D. J. Currie, 2007. The macroecological contribution to global change solutions. Science 316: 1581-1584.

Kottelat, M. & J. Freyhof, 2007. European freshwater fishes. Delemont, Switzerland.

Leprieur, F., O. Beauchard, S. Blanchet, T. Oberdorff & S. Brosse, 2008. Fish invasions in the world’s river systems: when natural processes are blurred by human activities. Public Library of Science, Biology 6(2): e28.

Leveque, C., T. Oberdorff, D. Paugy, M. L. J. Stiassny & P. A. Tedesco, 2008. Global diversity of fish (Pisces) in freshwater. Hydrobiologia. 595: 545–567.

Maitland, P.S., 2000. Guide to freshwater fish of Britain and Europe. Hamlyn ed., Octopus Pub Group. England.

MEA (Millennium Ecosystem Assessment), 2005. Ecosystems and human well-being: Biodiversity synthesis. Washington, DC: World Resources Institute.

Moyle, P.B. & J. J. Cech, 2004. Fishes: An introduction to ichthyology. New Jersey: Prentice-Hall. USA.

Oberdorff, T., J. F. Guegan & B. Hugueny, 1995. Global scale patterns of fish species richness in rivers. Ecography 18: 345–352.

Oberdorff, T., B. Hugueny & J. F. Guegan, 1997. Is there an influence of historical events on contemporary fish species richness in rivers? Comparisons betweenWestern Europe and North America. Journal of Biogeography 24: 461–467.

Oberdorff, T., P. A. Tedesco, B. Hugueny, F. Leprieur, O. Beauchard, S. Brosse & H. H. Dürr, 2011. Global and regional patterns in riverine fish species richness – A review. International Journal of Ecology doi: 10.1155/2011/967631.

Orme, C. D. L., R. G. Davies, M. Burgess, F. Eigenbrod, N. Pickup, V. A. Olson, A. J. Webster, T. S. Ding, P. C. Rasmussen, R. S. Ridgely, A. J. Stattersfield, P. M. Bennett, T. M. Blackburn, K. J. Gaston & I. P. F. Owens, 2005. Global hotspots of species richness are not congruent with endemism or threat. Nature. 436: 1016-1019.

Reyjol, Y., B. Hugueny, D. Pont, P.G. Bianco, U. Beier, N. Caiola, F. Casals, I. Cowx, A. Economou, T. Ferreira, G. Haidvogl, R. Noble, A. De Sostoa, T. Vigneron & T. Virbickas, 2007. Patterns in species richness and endemism of European freshwater fish. Global Ecology and Biogeography 16: 65–75.

Seys, J., P. Pissierssens, E. Vanden Berghe & J. Mees, 2004. Marine data management: we can do more, but can we do better? Ocean Challenge 13: 20–24.

Tedesco PA., T. Oberdorff, C.A. Lasso, M. Zapata & B. Hugueny, 2005. Evidence of history in explaining diversity patterns in tropical riverine fish. Journal of Biogeography 32: 1899-1907.

Tedesco, P. A., F. Leprieur, B. Hugueny, S. Brosse, H. H. Dürr, O. Beauchard, F. Busson & T. Oberdorff, 2012. Patterns and processes of global riverine fish endemism. Global Ecology and Biogeography (in press).

Villeger, S., S. Blanchet, O. Beauchard, T. Oberdorff & S. Brosse, 2011. Homogenization patterns of the world's freshwater fish faunas. Proceedings of the National Academy of Sciences of the United States of America. 108: 18003-18008.

Table 1. Overall content of the SPRICH database for each of the 6 biogeographic realms (Brealm). N basins is the number of basins. Values of species richness are mean number of species (± sd). Environmental descriptors are mean (± sd) altitudinal range from the source to the estuary expressed in metres, and mean (± sd) basin area in square kilometres.

		Fish species richness				Environmental descriptors	
B realm	N basins	Total	Native	Exotic	Endemic	Altitudinal range	Basin area
Afrotropical	108	59.33 (± 96.50)	57.88 (± 96.19)	1.45 (± 2.74)	13.69 (± 67.86)	913 (± 740)	186 214 (± 619 387)
Australian	179	17.01 (± 12.92)	14.74 (± 11.78)	2.26 (±2.76)	0.22 (± 0.87)	617 (± 452)	26 153 (± 121 089)
Nearctic	204	37.92 (± 40.44)	31.52 (± 35.32)	6.40 (± 10.06)	0.84 (± 4.00)	822 (± 815)	75 993 (± 297 085)
Neotropical	156	65.55 (± 177.52)	64.50 (± 177.84)	1.47 (± 2.58)	11.70 (±69.63)	1 405 (± 1 257)	97 543 (± 548 942)
Oriental	59	60.10 (± 108.99)	57.53 (± 105.11)	2.57 (± 4.79)	6.69 (±27.51)	1 936 (± 1 448)	91 005 (± 235 129)
Palearctic	348	25.12 (± 28.25)	21.09 (± 25.90)	4.02 (± 6.12)	0.86 (±8.35)	1008 (± 938)	81 764 (± 314 701)

Figure caption

Fig. 1. Map indicating the area covered by the 1054 river basins considered in the SPRICH database (in grey). Note that some large areas where no basin is informed account for deserts (both cold and hot) where there are no perennial rivers (e.g. northern Africa, central Australia, polar zones).

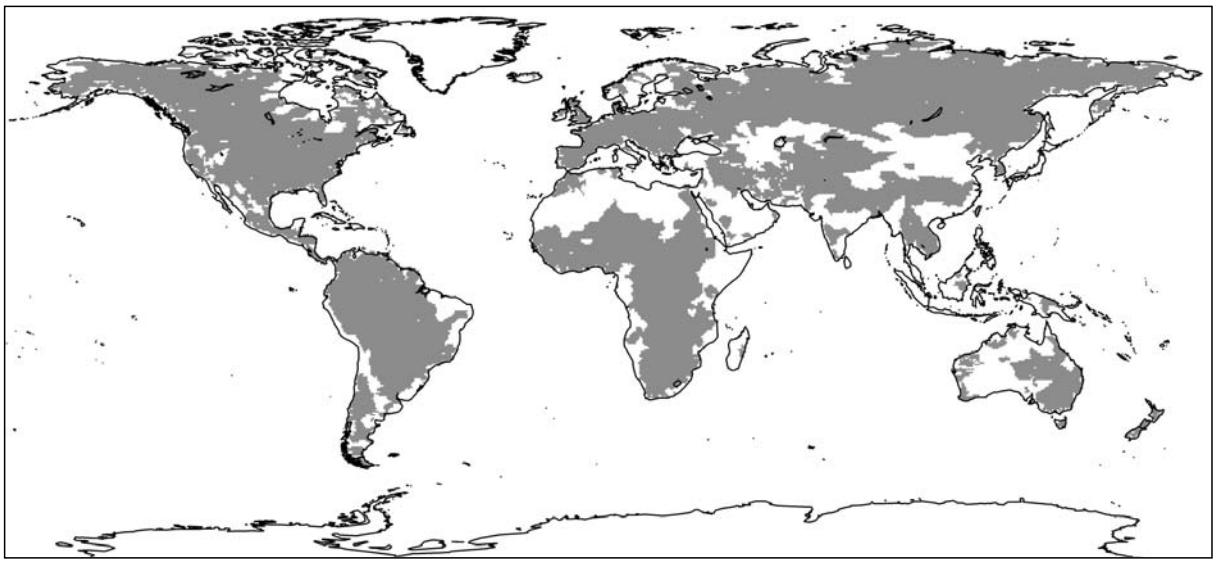


Figure 1

AUTHOR: Christine Lauzeral

TITLE: Using niche models to predict the invasive potential of non native species: methodological approaches and applications to freshwater fish on the French territory.

DIRECTOR: Sébastien Brosse

ABSTRACT:

Freshwaters are among the most anthropogenicaly threatened ecosystems in the world. They especially face serious threats due to invasive species. Efforts are thus needed to control invasions and increase the accuracy of the models used to predict the potential distribution of invasive species.

We showed that:

- the area of the observed distribution, and thus the area of the distribution predicted by correlative models, increases exponentially with the spatial grain of the data. However, model quality is little affected by the grain of the data and decreases only for the largest grains.
- coarse-grained occurrence data remain useful in identifying the species that experience niche shifts.
- most of the six fish species that we have studied were able to establish in France under climate change, even without niche shift.
- the iterative ensemble modeling method that we developed increases the accuracy of predictions as soon as the occurrence data set contained abundant non environmental absences. This new method is of interest for invasive species niche modeling but also to model the distribution of difficult to detect or endangered species.

KEY WORDS: non native species, niche models, correlative models, freshwater fish, niche shift, global change.

SUBJECT: Ecology

LABORATORY: Laboratoire Evolution & Diversité Biologique (EDB – UMR 5174), Université Paul Sabatier, 118 route de Narbonne, Bât 4R1, 31062 Toulouce cedex 9, France.

AUTEUR: Christine Lauzeral

TITRE: Prédiction du potentiel d'invasion des espèces non natives par des modèles de niche : approches méthodologiques et applications aux poissons d'eau douce sur le territoire français.

DIRECTEUR: Sébastien Brosse

LIEU ET DATE : Université Paul Sabatier, Toulouse, le 20 septembre 2012

RESUME :

Les espèces invasives sont une des perturbations les plus importantes des milieux aquatiques d'eau douce. La prévention des invasions passe par l'identification des zones dans lesquelles ces espèces sont susceptibles de s'installer.

Cette thèse a mis en évidence que :

- le changement de grain spatial modifie la mesure de l'aire de la distribution observée. Par contre, le grain des données utilisées pour calibrer les modèles influence peu la qualité des prédictions.
- les grains les plus larges restent utilisables pour identifier les espèces changeant de niche climatique lors du processus d'invasion.
- la majorité des six espèces de poissons étudiées présentent un fort risque d'établissement sous l'effet du changement climatique, même en l'absence de changement de niche de ces espèces.
- la qualité des prédictions des modèles corrélatifs peut être améliorée en utilisant la méthode d'ensemble itérative que nous avons développée. Cette méthode est particulièrement adaptée pour les espèces invasives, mais aussi pour les espèces difficiles à détecter ou les espèces en danger.

MOTS CLES : espèce non native, modèle de niche, modèle corrélatif, poissons d'eau douce, changement de niche, changement global.

DISCIPLINE : Ecologie

LABORATOIRE: Laboratoire Evolution & Diversité Biologique (EDB – UMR 5174), Université Paul Sabatier, 118 route de Narbonne, Bât 4R1, 31062 Toulouce cedex 9, France.